Le poids des mots

Volume II

Le poids des mots

Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles

Proceedings of the 7th International Conference on Textual Data Statistical Analysis

Louvain-la-Neuve
10-12 mars 2004 / March 10-12, 2004

Gérald Purnelle, Cédrick Fairon, Anne Dister (éds)

Volume II

UCL PRESSES UNIVERSITAIRES DE LOUVAIN

Comité de programme / Program Committee

Ramon Alvarez	Universitad de Leon, SP
Harald Baayen	Université de Nimègue, NL
Simona Balbi	Università di Napoli, IT
Valérie Beaudouin	France Télécom R&D, FR
Monica Bécue	Universitat Politècnica de Catalunya, SP
Vincent Blondel	Université catholique de Louvain, BE
Sergio Bolasco	Università degli studi di Roma La Sapienza, IT
Étienne Brunet	Université de Nice Sophia Antipolis, FR
Joseph Denooz	Université de Liège, BE
Jean-Claude Deroubaix	Université libre de Bruxelles, BE
Anne Dister	Université catholique de Louvain, BE
Michel Dorban	Université catholique de Louvain, BE
Annibale Elia	Università di Salerno, IT
Étienne Évrard	Université de Liège, BE
Cédrick Fairon	Université catholique de Louvain, BE
Serge Fleury	Université Paris 3 Sorbonne Nouvelle, FR
Corinne Gobin	FNRS, Université libre de Bruxelles, BE
Sylvianne Granger	Université catholique de Louvain, BE
Michel Kerboal	INSERM, Université de Rennes 1, FR
Dominique Labbé	Université de Grenoble, FR
Ludovic Lebart	CNRS, ENST Paris, FR (Président)
Alain Lelu	Université de Franche-Comté, FR
Sylvie Mellet	CNRS, Nice, FR
Piet Mertens	Katholieke Universiteit Leuven, BE
Annie Morin	IRISA, Université de Rennes 1, FR
Gérald Purnelle	Université de Liège, BE
Martin Rajman	EPFL Lausanne, CH
Max Reinert	CNRS, Université de Versailles SQY, FR
Marco Saerens	Université catholique de Louvain, BE
André Salem	Université Paris 3 Sorbonne Nouvelle, FR
Pascale Sebillot	IRISA, Université de Rennes 1, FR
Fiona Tweedie	University of Glasgow, RU
François Yvon	ENST, FR

Comité d'organisation / Organization Committee

Anne Dister	CENTAL, Université catholique de Louvain, BE
Cédrick Fairon	CENTAL, Université catholique de Louvain, BE
Gérald Purnelle	CIPL-LASLA, Université de Liège, BE

http://cental.fltr.ucl.ac.be http://www.ulg.ac.be/cipl/

Organisation locale / Local Organization

Bernadette Dehottay, Claude Devis, Michel Thomas, Laurent Simon, Patrick Watrin.
Conception du CD-Rom JADT 2004 : Parick Watrin.

© Presses universitaires de Louvain, 2004

Dépôt légal : D/2004/9964/11
ISBN : 2-930344-50-4

Imprimé en Belgique

Tous droits de reproduction, d'adaptation ou de traduction, par quelque procédé que ce soit, réservés pour tous pays, sauf autorisation de l'éditeur ou de ses ayants droit.

Diffusion :
www.i6doc.com, l'édition universitaire en ligne

Sur commande en librairie ou à :

Diffusion universitaire CIACO
Grand-Place, 7
1348 Louvain-la-Neuve, Belgique

Tél. 32 10 47 33 78
Fax 32 10 45 73 50
duc@ciaco.com

Table des matières / Table of Contents

Volume I

Keynote speakers

Douglas BIBER :
Conversation text types: A multi-dimensional analysis 15

Claudia LEACOCK :
Statistical Analysis of Text in Educational Measurement 35

Communications / Papers / Posters

Ramón ÁLVAREZ, Mónica BÉCUE, Olga VALENCIA :
Étude de la stabilité des valeurs propres de l'AFC d'un tableau lexical au moyen de procédures de rééchantillonage 42

Silvano AMATO, Emilio DI MEGLIO, Maria GUERRA :
Text Retrieval with External Information 52

Roxana ANGHELUTA, Patrick JEUNIAUX, Rudradeb MITRA, Marie-Francine MOENS :
Clustering Algorithms for Noun Phrase Coreference Resolution 60

Mappillairaju BAGAVANDAS, G. MANIMANNAM :
Quantification Of Stylistic Traits: A Statistical Approach 71

Simona BALBI, Emilio DI MEGLIO :
A Text Mining Strategy based on Local Contexts of Words 79

Ana-Maria BARBU :
Simple linguistic methods for improving a word alignment algorithm 88

Silvia BARTOLETTI, Alessandra GARBERO, Silvia MONTECOLLE, Ferdinando NISCO, Emanuela RECCHINI, Irene SALERNO :
Gli sbarchi dei clandestini nei quotidiani: un'analisi testuale esplorativa 99

Valérie BEAUDOUIN, François YVON :
Contribution de la métrique à la stylométrie 107

Mónica BÉCUE, Jérôme PAGÈS, Campo-Elias PARDO :
Analysis of multilingual free responses 119

Luc BÉLANGER, Guy LAPALME :
Identification de questions pour traiter les courriels par une méthode question-réponse 128

Jean-Guy BERGERON, Dominique LABBÉ :
Analyser les entretiens sociologiques 136

Charles BERNET :
Hasards de la rime 148

Anne BERRY, Bangaly KABA, Mohamed NADIF, Eric SANJUAN, Alain SIGAYRET :
Classification et désarticulation de graphes de termes 160

Yves BESTGEN :
Analyse sémantique latente et segmentation automatique de textes 171

Yves BESTGEN, Cédrick FAIRON, Laurent KERVES :
Un baromètre affectif effectif. Corpus de référence et méthode pour déterminer la valence affective de phrases 182

Ismaïl BISKRI, Jean-Guy MEUNIER, Sylvain JOYAL :
L'extraction des termes complexes : une approche modulaire semi-automatique 192

Sergio BOLASCO, Francesca DELLA RATTA – RINALDI :
Experiments on semantic categorisation of texts: analysis of positive and negative dimension 202

Mathieu BRUGIDOU, Nadine MANDRAN, Michel MOINE, Annie-Claude SALOMON :
Les apports de l'analyse textuelle pour l'analyse électorale : les questions ouvertes du panel électoral de 2002 211

Sylviane BURNER :
Le rapport à l'autre dans la psychose bipolaire 221

Carmela CAPPELLI, Angela D'ELIA :
La percezione della sinonimia: un'analisi statistica mediante moelli per ranghi 229

Simona CARBONE, Maria LONGOBARDI :
Gli aggettivi delle rappresentazioni di genere in adolescenza 241

Renzo CARLI, Francesca DOLCETTI, Nadia BATTISTI :
L'analisi emozionale del testo (AET): un caso di verifica nella formazione professionale 250

Antonio CHIRUMBOLO, Alessandra ARENI :
Linguaggio, ideologia e categorizzazione sociale: un'analisi psicologico sociale del documento di rivendicazione dell'attentato a Marco Biagi 262

Marie-Catherine de MARNEFFE, Pierre DUPONT :
Comparative study of statistical word sense discrimination techniques 270

Anne DE ROECK, Avik SARKAR, Paul H. GARTHWAITE :
Defeating the Homogeneity Assumption 282

Jean-Claude DEROUBAIX :
Que faire des corpus multilingues parallèles ? Une expérience 295

Guy DEVILLE, Laurence DUMORTIER, Hans PAULUSSEN :
Génération de corpus multilingues dans la mise en œuvre d'un outil en ligne d'aide à la lecture de textes en langue étrangère 304

Anne DISTER :
La féminisation des noms de métier, fonction, grade ou titre en Belgique francophone. État des lieux dans un corpus de presse 313

Hai DOAN-NGUYEN, Leila KOSSEIM :
Amélioration de la précision dans un système de question-réponse de domaine fermé 325

Antoine DOUCET :
Utilisation de séquences fréquentes maximales en recherche d'information 334

Patrick DROUIN :
Spécificités lexicales et acquisition de la terminologie 345

Jules DUCHASTEL, François DAOUST, Dimitri DELLA FAILLE :
SATO-XML : une plateforme Internet ouverte pour l'analyse de texte assistée par ordinateur 353

Jules DUCHASTEL, Francis J. LACOSTE, François PIZARRO NOËL :
Une stratégie intégrée de recherche en sciences humaines dans le Portail ATO-MCD 364

Anne DUFRESNE :
Le discours de la BCE concernant les aspects sociaux 373

Vincent J. DURIAU, Rhonda K. REGER :
Choice of Text Analysis Software in Organization Research: Insight from a Multi-dimensional Scaling (MDS) Analysis 382

Louisette EMIRKANIAN, Christophe FOUQUERÉ, Fabrice ISSAC :
Corpus issus du Web : analyse des pertinences thématique et informationnelle . 390

Frédéric ERLOS :
Référentiels terminologiques adaptables au contexte. L'exemple d'un système de recherche d'informations dans une grande entreprise 399

Stefan EVERT :
A simple LNRE Model for Random Character Sequences 411

Cédrick FAIRON, Ngoc-Diep HO :
Quantité d'information échangée : une nouvelle mesure de la similarité des mots 423

Dominic FOREST, Jean-Guy MEUNIER :
Classification et catégorisation automatiques : application à l'analyse thématique des données textuelles 434

François FOUSS, Jean-Michel RENDERS, Marco SAERENS :
Some relationships between Kleinberg's hubs and authorities, correspondence analysis, and the Salsa algorithm 445

Itsuko FUJIMURA, Mitsumi UCHIDA, Hiroshi NAKAO :
De *vs* des *devant les noms précédés d'épithète en français : le problème de* petit 456

Jean-Gabriel GANASCIA, Irène FENOGLIO, Jean-Louis LEBRAVE :
EDITE MEDITE : un logiciel de comparaison de versions 468

Claire GÉLINAS-CHEBAT, François DAOUST, Monique DUFRESNE, Karine GALLOPEL, Marie Élaine LEBEL :
Analyse exploratoire d'entrevues de groupe : les jeunes Français et le tabac 479

Raphaël GÉRARD, Bastien KINDT :
D'un dictionnaire de lemmatisation (D.A.G.) à un dictionnaire dérivationnel du grec ancien (D.D.G.) 488

Gaëtanelle GILQUIN, Eric LECOUTRE :
(How) can causative constructions be predicted? 496

Luca GIULIANO :
Il lessico della guerra nei newsgroup della categoria it.politica durante la guerra in Iraq 504

Cyril GOUTTE, Eric GAUSSIER, Nicola CANCEDDA, Hervé DÉJEAN :
Generative vs Discriminative Approaches to Entity Recognition from Label-Deficient Data 515

Maria Gabriella GRASSIA, Michelangelo MISURACA, Germana SCEPI :
Relazioni non Simmetriche tra Corpora 524

Edel P. GREEVY, Alan F. SMEATON :
Text Categorisation of Racist Texts Using a Support Vector Machine 533

Gaston GROSS :
Réflexions sur le traitement automatique des langues 545

Patricia GUILPIN, Christian GUILPIN :
Nouvelle méthode d'analyse statistique de la fréquence d'apparition d'un mot particulier (études synchroniques et diachroniques) 557

Benoît HABERT, Gabriel ILLOUZ, Helka FOLCH :
Dégrouper les sens : pourquoi, comment ? 565

Serge HEIDEN :
Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex 577

Fidelia IBEKWE-SANJUAN, Eric SANJUAN :
Mapping the structure of research topics through term variant clustering : the TermWatch system 589

Angel IGELMO, Gabriel M. JORDÀ, Carlota VICENS :
El análisis estadístico para el estudio de los campos estilísticos en una obra literaria 601

Volume II

Michel JACOBSON :
Corpus oraux glosés : outils logiciels d'aide à l'analyse 625

Bernard JACQUEMIN :
Analyse et expansion des textes en question-réponse 633

Jean-Marie JACQUES, Nathanaël LAURENT, Anne WALLEMACQ :
Paradoxes, dilemmes et contradictions : une mise en lumière au moyen du logiciel EVOQ© 644

Radwan JALAM, Jérémy CLECH, Ricco RAKOTOMALALA :
Cadre pour la catégorisation de textes multilingues 650

Michèle JARDINO :
Recherche de structures latentes dans des partitions de « textes » de 2 à k classes 661

Margareta KASTBERG SJÖBLOM :
Analyse grammatico-métrique d'une monographie "multi-générique" ; le substantif 672

Nicolas KUMPS, Pascal FRANCQ, Alain DELCHAMBRE :
Création d'un espace conceptuel par analyse de données contextuelles 682

Mathieu LAFOURCADE, Violaine PRINCE :
Modélisation de l'Hyperonymie via la combinaison de réseaux sémantiques et de vecteurs conceptuels 692

Anne-Catherine LANTIN, Philippe V. BARET, Caroline MACÉ :
Phylogenetic analysis of Gregory of Nazianzus' Homily 27 700

Ludovic LEBART :
Validité des visualisations de données textuelles 708

Jean-marc LEBLANC, Pierre FIALA :
Autour du Je présidentiel 716

Christophe LEJEUNE :
Représentation des réseaux de mots associés 726

Alain LELU :
Analyse en composantes locales et graphes de similarité entre textes 737

Dominique LONGRÉE, Xuan LONG, Sylvie MELLET :
Temps verbaux, axe syntagmatique, topologie textuelle : analyses d'un corpus lemmatisé 743

Jean-Luc MANGUIN :
L'évolution en français de l'adjectif épithète vers la postposition : réalité syntaxique ou trompe-l'œil lexical ? 753

Chantal-Édith MASSON, Hélène CAJOLET-LAGANIÈRE, Pierre MARTEL :
La BDTS-concordances : un outil technologique d'enrichissement de la pratique lexicographique 764

Denis MAUREL :
Les mots inconnus sont-ils des noms propres ? 776

Damon MAYAFFRE :
Analyse logométrique de la cohabitation Chirac/Jospin (1997-2002). Explication de la défaite de Lionel Jospin à l'élection présidentielle de 2002 785

Nicolas MAZZIOTTA :
Le texte dans tous ses états. Philosophie d'encodage du projet Khartês 793

Maura MISITI, Simona CARBONE :
Secondo gli esperti: Popolazione e società nelle opinioni dei testimoni privilegiati di una ricerca sugli adolescenti italiani 804

A. MOKRANE, R. AREZKI, G. DRAY, P. PONCELET :
Cartographie automatique du contenu d'un corpus de documents textuels 816

Rogério MUGNAINI, Esteban FERNANDEZ TUESTA, Adalberto OTRANTO TARDELLI :
Citations Titles Standardization Using Information Retrieval Techniques 824

Jean-Pierre MÜLLER :
ttda – une librairie R pour l'analyses de données textuelles 831

Takuya NAKAMURA :
Analyse automatique d'un discours spécialisé au moyen de grammaires locales 837

Berthille PALLAUD, Sandrine HENRY :
Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé 848

Sophie PIÉRARD, Liesbeth DEGAND, Yves BESTGEN :
Vers une recherche automatique des marqueurs de la segmentation du discours 859

Bénédicte PINCEMIN :
Lexicométrie sur corpus étiquetés 865

Carmen PINEIRA-TRESMONTANT :
Un pas en avant, un pas en arrière (Vingt-cinq ans d'allocutions radiodiffusées du roi Juan-Carlos d'Espagne) 874

Sophie PIRON :
Contraintes syntaxiques et préférences sélectionnelles du verbe entendre 885

Thierry POIBEAU :
Pré-analyse de corpus 897

Jean-Luc POMMIER :
Des variables tensives inscrites dans le texte : une interprétation dynamique de l'A.F.C. dans l'analyse d'Alceste 904

Yasmina QUATRAIN, Sylvaine NUGIER, Anne PERADOTTO, Damien GARROUSTE :
Évaluation d'outils de Text Mining : démarches et résultats 916

Paul RAYSON, Damon BERRIDGE, Brian FRANCIS :
Extending the Cochran rule for the comparison of word frequencies between corpora 926

Alex RIBA, Josep GINEBRA :
Diversity of Vocabulary and Homogeneity of Style in Tirant lo Blanc 937

Mathieu ROCHE, Thomas HEITZ, Oriane MATTE-TAILLIEZ, Yves KODRATOFF :
EXIT: un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés 946

Maria Clelia ROMANO, Tania CAPPADOZZI :
Il processo di codifica dei dati testuali dell'indagine Multiscopo "Uso del tempo" 958

Cristelle ROUX, Alain LEFÈVRE :
Identification des besoins en information géographique 970

Thibault ROY, Pierre BEUST :
ProxiDocs : un outil de cartographie et de catégorisation thématique de corpus 978

André SALEM :
Introduction à la résonance textuelle 987

Luiggi SANSONETTI :
Apports de la statistique textuelle pour le repérage des reprises et reformulations dans les corpus d'interaction verbale entre un adulte et un enfant 994

Jacques SAVOY, Yves RASOLOFO :
Hyperliens et recherche d'information sur le web 1001

Didier SCHWAB, Mathieu LAFOURCADE, Violaine PRINCE :
Hypothèses pour la construction et l'exploitation conjointe d'une base lexicale sémantique basée sur les vecteurs conceptuels 1009

Gilda SENSALES, Antonio CHIRUMBOLO :
Le rappresentazioni delle differenze di "Genere" nel mondo del lavoro attraverso la comunicazione scientifica in psicologia sociale: analisi del lessico degli "Psychological Abstracts" (1976-2002) 1020

Benedikt SZMRECSÁNYI :
On Operationalizing Syntactic Complexity 1032

Cristina Alice TOMA :
Cohésion informative dans le discours scientifique 1040

Carlo TOMASETTO, Patrizia SELLERI :
Lessico dell'intervista, lessico degli intervistati : l'articolazione tra domande e risposte nell'analisi di Alceste 1052

Stéphane TRÉBUCQ :
Finance organisationnelle : un essai de représentation 1062

Laurence TUERLINCKX :
La lemmatisation de l'arabe non classique 1070

Jose TUMMERS, Dirk SPEELMA, Dirk GEERAERTS :
Quantifying semantic effects. The impact of lexical collocations on the inflectional variation of Dutch attributive adjectives 1080

Gian Piero TURCHI, Sara MARTINAZIOLI, Luisa ORRÙ, Barbara LALISCIA :
La "malattia mentale" tra senso scientifico e senso comune: analisi di testi di settore e divulgativi 1090

Arjuna TUZZI, Marisa CEMIN, Marco CASTAGNA :
"Moved deeply I am". Autistic language in texts produced with FC 1098

Mathieu VALETTE, Natalia GRABAR :
Caractérisation de textes à contenus idéologiques : statistique textuelle ou extraction de syntagme ? l'exemple du projet PRINCIP 1107

Valery VANDAELE, Pascal FRANCQ, Alain DELCHAMBRE :
Analyse d'hyperliens en vue d'une meilleure description des profils 1118

Lieve VANGEHUCHTEN :
El uso de la estadística en la didáctica de las lenguas extranjeras con fines específicos: descripción del proceso de selección del léxico típico del discurso económico empresarial en español 1129

Sofie VAN GIJSEL, Dirk GEERAERTS, Dirk SPEELMAN :
A functional analysis of the linguistic variation in Flemish spoken commercials 1037

Fabienne VENANT :
Polysémie et calcul du sens 1146

Jacques VERGNE :

Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource 1158

Jean-Marie VIPREY :

Analyse séquencée de la micro-distribution lexicale 1166

Hung VO TRUNG :

SANDOH, un outil pour analyser des textes hétérogènes 1178

David WARTEL, Pascal FRANCQ, Alain DELCHAMBRE :

Organisation d'une masse documentaire électronique présentée à des lecteurs potentiels 1186

Maria ZIMINA :

Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles 1196

Tables rondes / Workshops

Lexicométrie et corpus multilingues 1204

Corneille et Molière 1208

Index

Index des auteurs /Authors Index 1209

Index des mots-clés / Keywords Index 1213

Corpus oraux glosés : outils logiciels d'aide à l'analyse

Michel Jacobson

LACITO - CNRS – 7, rue Guy Môquet, Bât 23 – 94801 Villejuif Cedex – France
jacobson@idf.ext.jussieu.fr

Abstract

Field linguistics works on written data, usually created through the analysis of oral recordings. The structure and the type of searches one may desire can be specific to aspects of sound or temporality. Apart from these few characteristics, the methods used to study this type of data have many points in common with those used in textual linguistics.
To illustrate these common points, we will examine several cases of recurrent tasks in field linguistics: morphemic analysis and glossing. We will see how recourse to the study of segments contexts (words, morphemes) given by concordances, or recourse to the calculation of the frequency of use of these units can help to carry out these tasks, or even partially automate them.
To finish, we will present a program designed to help gloss oral data by compiling, alongside the data, a lexicon of all the glosses already used. We will also examine how to use this tool to implement the methods cited above to optimize its usefulness for linguists.

Résumé

La linguistique de terrain travaille sur des corpus écrits, en général issus d'analyses d'enregistrements oraux. La structure et le type de requêtes que l'on peut formuler sur ces corpus peuvent être spécifiques de l'aspect sonore ou temporel. Mis à part ces caractéristiques, les méthodes utilisées pour l'étude de ces corpus ont de nombreux points en commun avec celles utilisées en linguistique textuelle.
Pour illustrer ces points communs, nous examinerons quelques cas de tâches récurrentes en linguistique de terrain : la segmentation et la détermination de gloses. Nous verrons notamment comment le recours à l'examen des contextes d'apparition de segments (mots, morphèmes) donnés par des concordances, ou bien comment le calcul des fréquences d'apparition de ces unités peuvent nous aider à effectuer ces tâches ou bien même à les automatiser en partie.
Nous présenterons enfin un logiciel créé pour aider à gloser des corpus oraux en entretenant parallèlement au corpus, un lexique de toutes les gloses déjà utilisées. Nous examinerons également comment implémenter dans cet outil les méthodes citées plus haut pour optimiser l'aide apportée au linguiste.

Mots-clés : linguistique de terrain, corpus oraux, langage de balisage de texte.

1. Présentation

Beaucoup de laboratoires de recherche en linguistique, dont le LACITO[1] auquel nous appartenons, détiennent et gèrent des matériaux récoltés lors d'enquêtes de terrain, notamment des enregistrements de parole (récits, dialogues, questionnaires, etc.). Ces enregistrements ont fait, pour partie, l'objet d'analyses comportant la plupart du temps une transcription « phonétique », mais aussi parfois des traductions, des gloses, des indications scénographiques, etc. L'ensemble de ces données (enregistrements et analyses) pour une langue donnée constitue un

[1] Laboratoire « LAngues et CIvilisations à Tradition Orale » du CNRS. Le LACITO est un laboratoire dont les chercheurs (linguistes, anthropologues et ethno-musicologues) travaillent depuis plus de 30 ans à la constitution d'une documentation des traditions orales un peu partout dans le monde.

corpus sur lequel les chercheurs vont pouvoir continuer à travailler après leur mission sur le terrain.

Les enregistrements ont été stockés suivant les époques, sur des supports (mécaniques, magnétiques, optiques) analogiques puis plus récemment numériques. Les analyses ont, elles, été stockées parfois sous forme manuscrite, d'autres fois sous forme imprimée et pour les plus récentes d'entre elles, il peut exister une forme électronique informatique dans des formats divers : traitement de texte, base de données ou texte structuré de type Shoebox (Buseman. et Buseman, 1998), Lexware (Hsu, 1989), etc.

1.1. *Les spécificités de la linguistique de terrain*

Ces données (enregistrements et analyses) représentent souvent la rare documentation dont disposent les linguistes sur leurs langues. Ces langues étant souvent minoritaires, peu étudiées et parfois en danger de disparition ou même déjà disparues, les quelques rares données collectées, constituent à la fois un patrimoine culturel irremplaçable et du matériel précieux pour la recherche.

Le volume et la nature même des données disponibles ne nous permettent pas d'utiliser tous les outils que l'on pourrait s'attendre à trouver pour des langues bien documentées comme l'anglais ou le français. Au mieux, nous disposons pour une langue donnée : d'un corpus de quelques heures d'enregistrements accompagnés d'une description au niveau morpho-phonologique, quelquefois d'un lexique ou d'un dictionnaire de quelques milliers de mots, et éventuellement d'une ébauche de grammaire. Les enregistrements dont nous disposons sont assez éloignés des conditions de laboratoire en chambre sourde, d'une part parce qu'ils ont été pratiqués sur le terrain en situations naturelles (récits avec un conteur et un public, cérémonies, dialogues dans la rue), d'autre part parce que le matériel d'enregistrement utilisé devait être aussi léger et autonome en énergie que possible. Il n'est pas possible non plus, étant donné le peu de personnes travaillant sur ces langues de trouver sur le marché des outils de type analyseur syntaxique, ou lemmatiseur… La lourdeur de mise en œuvre de certains de ces outils constitue souvent un obstacle pour des utilisations ponctuelles ou pour des trop petits volumes de données, et rend les interventions manuelles compétitives en termes d'effort et d'investissement en temps.

1.2. *Les méthodes utilisées*

L'examen des données d'un corpus emprunte en partie ses méthodes aux disciplines du traitement du signal. En particulier, l'examen d'un enregistrement oral peut aujourd'hui se faire facilement avec des outils logiciels appliquant des techniques DSP[2] pour présenter le signal de parole sous forme de spectrogrammes, pour faire des analyses formantiques, des extractions de fondamentale, etc. Toutes ces techniques concourent pour le linguiste à se faire une meilleure opinion des contraintes phonético-phonologiques et donc à améliorer son système de transcription.

Une fois établie une première transcription, considérée comme une analyse et non comme une donnée, la plupart des méthodes utilisées par la suite pour questionner, structurer ou enrichir cette annotation sont inspirées des traditions d'études textuelles. En particulier, les linguistes font un usage intensif de concordances. Celles-ci permettent de se faire une meilleure idée des conditionnements qui peuvent exister entre les segments au niveau phonologique. Elles sont aussi utilisées par les linguistes pour distinguer les différents usages d'une forme lexicale afin

[2] Digital Signal Processing.

d'en extraire les différents sens pour la constitution d'un dictionnaire. Un autre outil simple est l'utilisation des fréquences d'apparition de certaines unités, ceci afin de repérer les formes rares, déviantes, voir même les erreurs de transcription.

Dans ce qui suit, nous allons présenter deux types d'activités habituelles aux linguistes de terrain (la segmentation et la détermination de gloses) dans lesquelles les méthodes de travail sur corpus que nous venons de citer sont utilisées en tant qu'aide.

2. Les aides à la segmentation

La segmentation est une opération récurrente que le linguiste utilise à différentes étapes de son travail d'analyse. La transcription « phonétique » telle que la pratique les linguistes, en utilisant des systèmes de notation alphabétiques de type API (Alphabet Phonétique International), présuppose déjà une segmentation implicite de la parole en unités discrètes. D'autres étapes de segmentation en unités plus larges telles que les morphèmes, les mots ou les syntagmes, s'aident d'analyses pratiquées sur des extraits de corpus. Par exemple, on peut facilement concevoir un logiciel d'aide au découpage morphologique, lequel partant d'une transcription d'un mot, proposerait en priorité les segmentations déjà rencontrées, voire les plus fréquentes ou les plus probables.

La segmentation par rétroconversion

Les transcriptions sur lesquelles nous travaillons d'habitude, marquent les unités lexicales et morphologiques au moyen de conventions typographiques. Il est beaucoup plus rare de rencontrer des transcriptions « au kilomètre ». Parfois, le découpage morphologique n'est pas indiqué ou ne l'est pas systématiquement. Cela correspond en général au choix de l'analyste de ne pas pousser son analyse plus loin, par exemple en raison de la difficulté de la tâche (présence de nombreux amalgames, etc.). La segmentation revient donc, la plupart du temps, à utiliser ces indices typographiques pour isoler et étiqueter de manière explicite les segments. C'est ce que l'on appelle la rétroconversion.

La « rétroconversion » est une tâche qui consiste à retrouver la structure logique sous-jacente ayant donné lieu à une structure physique particulière. Suivant les moyens utilisés pour le codage de l'analyse, cette étape de rétroconversion sera ou non complexe et nécessaire. Par exemple, dans les cas que nous traitons dans notre laboratoire, une grosse partie des analyses ayant été faite depuis assez longtemps, parfois avant la popularisation de la micro-informatique, la structure physique est une structure typographique destinée à l'impression sur un support papier. Heureusement, ces documents utilisent la plupart du temps une partie relativement standardisée de conventions de notation, comme par exemple l'utilisation des espaces blancs pour séparer les mots ou bien celle des tirets ou des points pour indiquer les jointures de morphèmes. La rétroconversion peut dans ce cas être effectuée par une simple segmentation basée sur des séparateurs typographiques, et être réalisée par des jeux d'expressions régulières.

Beaucoup de nos documents d'analyse utilisent ces conventions dans un style de présentation « interlinéaire » assez habituel pour les linguistes. Dans un texte interlinéaire, les unités d'analyses mots se suivent traînant avec eux sur plusieurs lignes toute leur analyse: transcription(s), glose(s), etc. La ligne typographique est donc utilisée à la fois pour noter la succession des unités mais aussi les différentes analyses linguistiques de ces unités.

La gestion de l'alignement interlinéaire, suivant le système utilisé pour sa visualisation, peut être effectuée soit par des calculs sur la taille en caractères des unités (Buseman et Buseman,

1998), soit par des structures de données plus abstraites comme les tableaux HTML, ou des boites LaTex, soit enfin par des calculs sur de simples séparateurs de champs comme les caractères de tabulation. La rétroconversion, c'est-à-dire la redécouverte de la segmentation sous-jacente sera, suivant les cas, plus ou moins aisée. Elle peut être effectuée en général, par des jeux d'expressions régulières mais demande parfois quelques interventions manuelles.

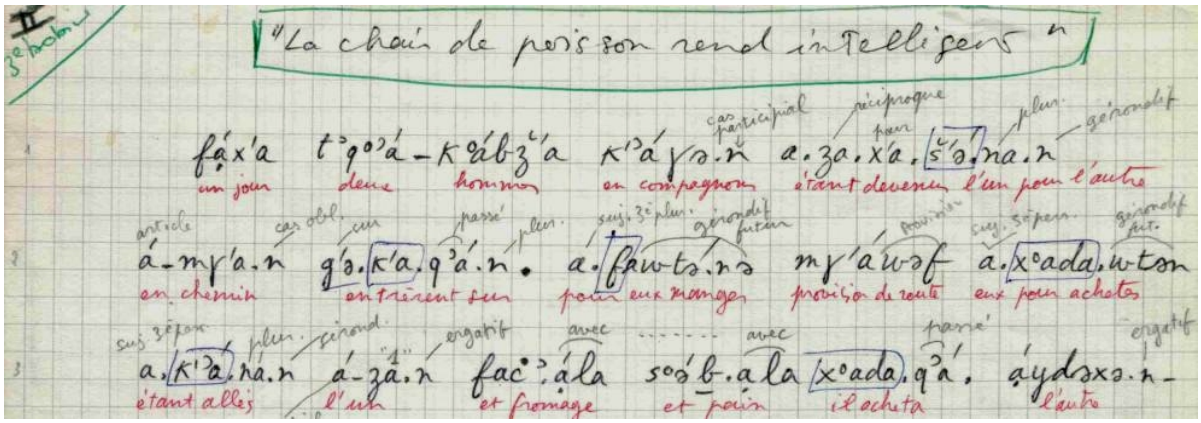


Figure 1. Exemple d'un texte[3] manuscrit dans une présentation « interlinéaire » [extrait d'un cahier manuscrit attribué à Georges Dumézil].

En règle générale, les solutions utilisées restent *ad hoc*. Les techniques employées en Traitement du Langage Naturel, à base d'automates à états finis (Kaplan et Martin, 1995), de modèles Markoviens (Rabiner, 1989), de règles sur des connaissances, ou bien à base de systèmes mixtes (Spoart *et al.*, 1996), sont souvent plus lourdes à mettre en œuvre qu'une correction manuelle ou qu'une re-saisie complète des données, à moins d'avoir à traiter de gros volumes de données ce qui est rarement le cas pour les langues que nous étudions.

3. Les aides à la détermination des gloses d'un texte

Une autre tâche assez récurrente dans la constitution de corpus oraux est la détermination des gloses des unités. Par cette opération, nous entendons l'action d'associer un commentaire à une unité d'analyse dans le but d'apporter une explication ou une définition de cette unité. Pour des niveaux d'analyse assez larges, comme le texte ou la phrase, on parlera plus facilement de traduction. Pour des niveaux plus bas d'analyse, comme le mot ou le morphème, on parlera plus volontiers de glose. Cette glose peut être plus ou moins technique, normalisée et en inventaire fini. Cette opération est délicate dans la mesure où elle nécessite une bonne connaissance de la langue en question ainsi qu'une rigueur et un systématisme. Elle est aussi une tâche extrêmement répétitive et fastidieuse car en règle générale il existe dans toute les langues un petit ensemble d'unités très souvent utilisées, et qu'il faut donc gloser de nombreuses fois.

Il s'agit donc d'une situation idéale pour proposer une aide de type logiciel. Un logiciel pourrait permettre à un utilisateur de choisir ses gloses dans un ensemble fini de gloses, ce qui lui permetrait d'être plus facilement systématique et d'éviter de gloser une même unité par des variantes de type : Passé, PASSE. Past, … Un logiciel pourrait aussi permettre de proposer en fonction de critères comme la forme de l'unité, la glose qui lui correspond s'il n'y en a

[3] Texte oubikh (langue du Caucase, aujourd'hui disparue). Il est possible de consulter ce conte (enregistrement, transcription et traduction) sur le site web du LACITO.

qu'une, ou les gloses possibles s'il y a de l'homophonie. Il pourrait aussi pour les cas d'homophonie, trier les propositions afin de faire apparaître en premier les plus probables.

C'est principalement pour des raisons économiques (gain de temps) que nous avons envisagé dans notre laboratoire de construire un outil logiciel d'aide au linguiste pour cette tâche.

3.1. *Description du logiciel*

Le logiciel créé (Interlinear Text Editor[4]) permet de présenter, pour un même document XML (Bary *et al.*, 1998), quatre niveaux d'analyse distincts : le texte, les phrases, les mots et les morphèmes. Il offre pour chacun de ces niveaux une représentation adéquate. Le mode de représentation d'un texte ou d'une phrase permet la visualisation et l'édition de sa transcription et de sa traduction. Le mode de représentation des mots et des morphèmes se fait phrase par phrase sous forme interlinéaire (cf. Figure 2). La glose des mots et des morphèmes est affichée dans un menu déroulant qui permet de changer sa valeur en la choisissant dans l'ensemble de toutes les autres gloses du même document ou des autres documents ouverts.

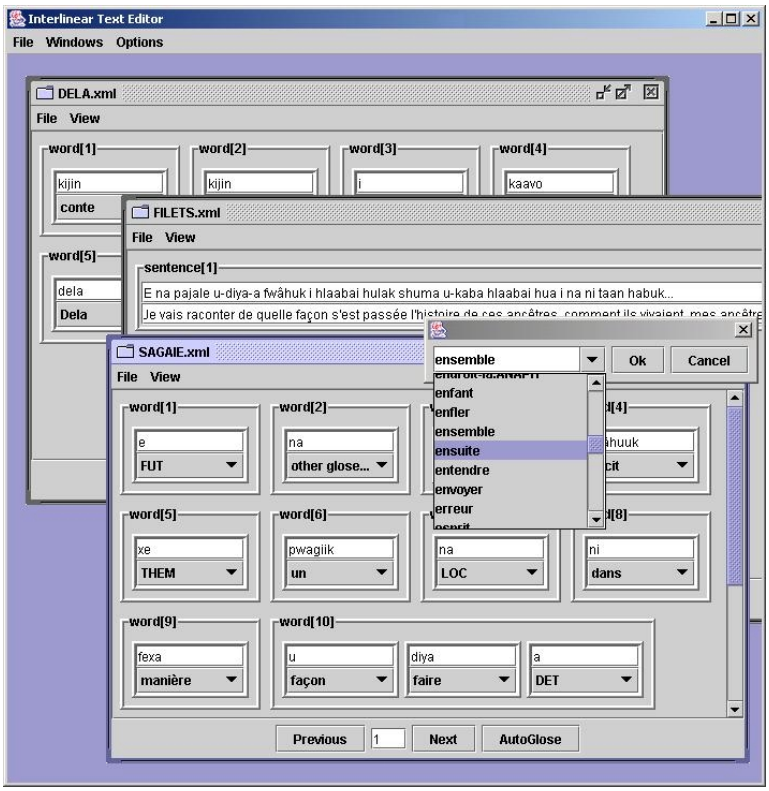


Figure 2. Logiciel « Interlinear Text Editor »

Le logiciel entretient une relation permanente et étroite entre les documents ouverts et un lexique construit de toutes les gloses utilisées dans ces derniers. Chaque transcription de mot ou de morphème est liée à une liste de gloses déjà utilisées pour les gloser. Cette liste comporte pour l'instant uniquement le contenu textuel des gloses et le nombre de fois où elles ont été utilisées. Tout ajout ou changement d'une glose dans un document entraîne un change-

[4] Le logiciel 'Interlinear Text Editor' écrit par l'auteur (Michel Jacobson) est librement disponible sur le site web (http://michel.jacobson.free.fr). Il s'agit d'un programme écrit en Java utilisable sur n'importe quelle plateforme disposant d'une machine virtuelle JAVA1.3.

ment dans ce lexique pour ajouter ou supprimer un lien ou bien augmenter ou diminuer le nombre d'occurrences de ce lien.

Il est possible aussi d'importer en bloc tout un lexique comportant l'ensemble des liens transcription/glose. Ce lexique sera simplement ajouté à celui calculé à partir des fichiers ouverts.

3.2. *Utilisation du logiciel*

À titre d'exemple, nous avons testé ce logiciel sur un corpus en langue nêlêmwa de Nouvelle-Calédonie. Nous disposons pour cette langue de quelques ressources :

- Un corpus de cinq textes narratifs (Lacito, 2003), représentant environ 1 heure 40 minutes d'enregistrement, accompagné d'une analyse explicitée en XML en phrases, en mots et en morphèmes, avec pour les phrases une transcription phonétique, une traduction en français ou en anglais, et des gloses en français pour chacun des morphèmes ;
- Un corpus d'une trentaine de textes formatés avec Shoebox, représentant 5 heures d'enregistrement. Chaque texte est analysé en phrases comportant une transcription phonétique et une traduction libre en français. Les mots et morphèmes sont indiqués dans la transcription par des marqueurs typographiques. Les morphèmes ne sont pas glosés ;
- Un dictionnaire formaté avec Shoebox comportant près de 3500 entrées (Bril, 2000);

Ceci est un exemple assez représentatif des ressources informatiques dont nous pouvons disposer dans notre laboratoire (et plus généralement en linguistique de terrain) pour une langue donnée.

Dans un premier temps nous avons construit un lexique à partir de 4 textes déjà glosés (environ 700 items), afin de disposer d'une source unique et synthétique d'informations.

Puis nous avons utilisé le logiciel pour gloser un cinquième texte. Nous avons compté :

1) le nombre de fois où le logiciel ne trouve qu'une manière de gloser un morphème. Ce dernier peut donc être glosé sans intervention de l'utilisateur. Nous avons distingué : 1a) les cas où le morphème est correctement glosé ; 2b) des cas où il ne l'est pas ;
2) le nombre de fois où le logiciel propose à l'utilisateur de choisir la glose parmi une liste de gloses. Nous avons distingué : 2a) les cas où la glose correcte faisait partie de la liste en première position, 2b) les cas où elle se trouvait dans les positions suivantes et 2c) les cas où elle n'en faisait pas partie.
3) le nombre d'échecs, c'est-à-dire qu'aucune glose n'est proposée par le logiciel ;

L'hypothèse que nous formulons est que le logiciel sera d'autant plus utile que le nombre de gloses correctement trouvées sera important et le nombre de gloses erronées faible. D'autre part le gain de temps sera d'autant plus grand que la glose correcte proposée en cas d'homophonie sera placée souvent en première position dans la liste. Enfin le logiciel devrait être capable de s'améliorer sur ces deux points par apprentissage. Cet apprentissage consiste à enrichir le lexique au fur et à mesure que le logiciel glose les morphèmes et à corriger à la volée ces choix si ceux-ci sont erronés. Pour simuler cet apprentissage nous avons utilisé le logiciel pour gloser un texte préalablement glosé manuellement. Le logiciel peut alors simuler les corrections d'un utilisateur en contrôlant et rectifiant automatiquement ses choix par comparaison avec les gloses préexistantes.

3.3. *Analyse des résultats*

	Correctement glosé (1a)	Incorrectement glosé (1b)	Total
Sans apprentissage	130	33	163
Avec apprentissage	158	19	177

Tableau 1.
Nombre de morphèmes correctement et incorrectement glosés en fonction de l'apprentissage.

Ce premier tableau (Tableau 1) ne comptabilise que les transcriptions de morphèmes présentes dans le lexique et ne présentant pas d'homophonie. Une seule glose est donc proposée par le logiciel. Elle est jugée correcte ou non en fonction de la glose effectivement présente dans le document d'origine. La différence dans le total des deux situations (avec et sans apprentissage) vient du fait que certains cas d'homophonie ont pu être détecté par l'apprentissage. On remarquera dans le tableau que globalement on obtient plus de bonnes propositions que de mauvaises et que ce taux s'améliore avec l'apprentissage (on passe de 81,6% à 89,2% de gloses correctes).

En examinant de manière qualitative les résultats on peut voir qu'un certain nombre de mauvaises propositions est du à l'utilisation de synonymes pour les gloses. Par exemple avec le corpus testé, la transcription [hulak] laquelle est glosée « ancien » dans le dictionnaire a été glosée « vieux » dans le corpus, alors que la transcription [hulaxa] laquelle, inversement, a été glosée « vieux » dans le dictionnaire à été cette fois glosée « ancien » dans le corpus. Un autre morphème précédemment glosé « semblable » a été cette fois glosé par « être comme ». [kûûlâ] qui avait été jusqu'alors glosé soit par « accomplir » soit par « finir », l'est cette fois dans le corpus par « terminer ». Nous pourrions citer de nombreux autres exemples.

	Position dans la liste			
	Première (2a)	Suivantes (2b)	Absente (2c)	Total
Sans apprentissage	162	194	73	429
Avec apprentissage	164	269	28	461

Tableau 2. Nombre de fois où la glose correcte apparaît dans la liste de propositions, en fonction de sa position dans la liste et de l'apprentissage.

Le tableau 2 comptabilise les transcriptions de morphèmes présentes dans le lexique et qui présentent de l'homophonie. Le logiciel calcule alors la liste de toutes les gloses candidates qu'il trie par fréquence d'apparition. On observe que le nombre de cas d'homophonie non reconnue diminue assez fortement (col. 2c : 28 cas contre 73 cas sans apprentissage), ce qui est logique pour un système avec apprentissage. L'apprentissage semble donc profiter en grande partie aux homophones. Mais le critère de fréquence d'utilisation utilisé ici ne permet pas d'améliorer le taux de bon classement de ces derniers (col. 2a et 2b). Ces résultats semblent nous encourager dans la voie de l'utilisation d'un autre critère tel que celui de la co-occurrence mise en valeur dans les concordances. Ce dernier critère devrait améliorer le classement de gloses des homophones en faisant apparaître en premier les plus probables en termes d'environnement contextuel et non plus en termes de simple fréquence d'utilisation.

	Aucune glose proposée (3)
Sans apprentissage	96
Avec apprentissage	50

Tableau 3. Nombre d'échecs en fonction de l'apprentissage

Le dernier tableau montre simplement que l'apprentissage a permis d'ajouter une cinquantaine d'entrées au lexique. L'examen du tableau 2 nous a déjà montré que ces nouvelles entrées ajoutent principalement de nouveaux cas d'homophonie.

4. Perspectives

Segmenter et gloser la transcription d'un texte oral sont deux tâches à la fois répétitives, fastidieuses et délicates. Un outil logiciel ayant pour objectif d'apporter une aide à ces deux tâches doit utiliser au maximum le peu de renseignements qu'il détient sur la langue pour que ses propositions d'analyses soient les plus fiables et pertinentes possibles. Etant donné le faible niveau de renseignements que l'on possède en général sur ces langues, le système d'aide doit pouvoir s'enrichir au fur et à mesure de l'analyse. C'est ce que nous avons tenté de faire avec le logiciel présenté plus haut. Il reste tout de même de nombreuses fonctions à ajouter à celui-ci pour qu'il puisse constituer une véritable aide au linguiste. De nombreuses interfaces restent aussi à définir pour qu'il soit plus convivial dans son utilisation et plus paramétrable afin de l'adapter à des besoins spécifiques divers.

Il convient aussi d'optimiser certains algorithmes pour que l'aide apportée soit plus pertinente. Le choix et l'ordre de présentation des gloses proposées pourrait venir par exemple de l'utilisation des concordances, ou bien de critères supplémentaire à la transcription, comme par exemple de connaissances syntaxiques (partie du discours, etc.).

Références

Buseman A et Buseman K. (1998). *The Linguists SHOEBOX*. Summer Institute of Linguistics. (http://www.sil.org/computing/catalog/shoebox.html).

Bray T., Paoli J. et Sperberg-McQueen C.M. (Eds) (1998). *Extensible Markup Language (XML) 1.0.* W3C Recommendation, 10 February 1998 (http://www.w3.org/TR/REC-xml).

Bril I. (2000). *Dictionnaire nêlêmwa-nixumwak-français-anglais*. LCP 14. Peeters.

Hsu R. (1989). *Lexware Manual*. Second Edition. Linguistics Dept., University of Hawaii.

Kaplan M.R. et Martin K. (1994). *Regular Models of Phonological Rule Systems*. Computational Linguistics, vol. (20/3) : 331-378.

Lacito. (2003). Site web du programme "Archivage" (http://lacito.vjf.cnrs.fr/archives).

Rabiner R.L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, vol (77/2) : 257-286.

Sproat R. *et al.* (1996). A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. In *Proceedings of the Meetng of the Association for Computational Linguistics* : 377-404.

Analyse et expansion des textes en question-réponse

Bernard Jacquemin

ILPGA – 19, rue des Bernardins – 75 005 Paris – France
b_jacquemin@yahoo.fr

Abstract

This paper presents an original methodology to consider question answering. We noticed that query expansion is often incorrect because of a bad understanding of the question. But the automatic good understanding of an utterance is linked to the context length, and the question are often short. This methodology proposes to analyse the documents and to construct an informative structure from the results of the analysis and from a semantic text expansion. The linguistic analysis identifies words (tokenization and morphological analysis), links between words (syntactic analysis) and word sense (semantic disambiguation). The text expansion adds to each word the synonyms matching its sense and replaces the words in the utterances by derivatives, modifying the syntactic schema if necessary. In this way, whatever enrichment may be, the text keeps the same meaning, but each piece of information matches many realisations. The questioning method consists in constructing a local informative structure without enrichment, and matches it with the documentary structure. If a sentence in the informative structure matches the question structure, this sentence is the answer to the question.

Résumé

Cet article présente une méthode originale d'envisager la tâche de question-réponse. Nous avons remarqué que l'expansion de requête est souvent erronée du fait d'une mécompréhension de la question. Mais la bonne compréhension d'un énoncé est fonction de la taille du contexte, et les question sont souvent courtes. Notre approche propose d'analyser les documents et de construire une structure informationnelle en utilisant les résultats de l'analyse ainsi que l'enrichissement sémantique des textes. Une analyse linguistique identifie les mots (segmentation et analyse morphologique), les liens entre les mots (analyse syntaxique) et le sens des mots (désambiguïsation sémantique lexicale). L'expansion de texte adjoint à chaque mot les synonymes qui correspondent à son acception contextuelle et remplacent les mots des énoncés par leurs dérivés, modifiant si nécessaire la structure syntaxique de la phrase. De la sorte, quel que soit l'enrichissement utilisé, le texte conserve la même signification, tandis que chaque élément d'information obtient de nombreuses actualisations. La procédure d'interrogation consiste à construire une structure informationnelle locale à la question sans effectuer d'enrichissement, et à faire lui faire correspondre un ou plusieurs fragments de la structure documentaire. Si une phrase de la structure informationnelle documentaire correspond à celle de la question, cette phrase contient la réponse à la question.

Keywords: question answering, information structuring, lexical semantic disambiguation, parsing, text expansion, synonymy, derivation.

1. Introduction

Dans notre société, la maîtrise de l'information est devenue un enjeu très important, que ce soit en politique, en économie, dans le domaine culturel, etc. Mais le contrôle de l'information doit faire face à la profusion des données disponibles. De fait, le nombre des documents textuels augmente chaque jour, et aucun humain ne peut actuellement avoir une idée précise de l'ensemble de l'information qu'ils contiennent.

De ce fait, les méthodes qui permettent d'étudier, d'identifier et de classifier l'information dans les textes électroniques prennent de plus en plus d'importance. Les disciplines de traitement de

l'information - et plus particulièrement l'extraction d'information, la recherche d'information ou la tâche de question-réponse - permettent de trouver un élément d'information dans une base textuelle. Dans ce domaine, chacune des disciplines possède ses propres spécificités. En particulier, la tâche de question-réponse joue un rôle de premier plan d'une part parce que contrairement à la recherche d'information, c'est une information particulière et précise qui doit être obtenue, et d'autre part parce que les données recherchées sont variable, contrairement à la discipline d'extraction d'information.

Depuis TREC-8 (*Text REtrieval Conference*) (Voorhees et Harman, 1999 ; Voorhees, 1999), toutes les méthodes de question-réponse proposées utilisent une architecture similaire : un module d'analyse de la question chargé d'identifier les éléments d'information présents dans la requête, et parfois de catégoriser la réponse attendue ; un système d'expansion de requêtes qui donne à chaque élément d'information de la question autant d'actualisations qu'il est possible, car la même donnée peut se présenter sous plusieurs formes dans un texte ; un moteur de recherche qui recherche chaque élément d'information dans les textes, quelle que soit son actualisation.

Toutes les méthodes qui participent à la compétition TREC fondent leur méthode de recherche sur le traitement de la requête. Par exemple, Hull (1999) utilise des analyseurs pour extraire et identifier le vocabulaire de la question et trouver la réponse contenant les mots les plus importants de cette question. Le système QALC (Ferret *et al.*, 1999) est basé sur une analyse linguistique de la requête, ainsi que sur l'exploitation par un moteur de recherche de mots-clefs obtenus au départ de cette requête (racinisation, synonymes, entités nommées etc.). Falcon (Harabagiu *et al.*, 2002) est le seul système qui considère que le traitement des données textuelles est aussi important que la catégorisation des questions. Cependant, ce système est également centré sur le traitement de la requête. En effet, les fragments de la base textuelle analysés sont obtenus par un moteur de recherche utilisant le vocabulaire de la question ou des mots-clefs issus du traitement de cette question.

Nous avons décidé de faire varier cette architecture traditionnelle en effectuant les principaux traitements non plus sur la requête, mais sur les textes de la base documentaire dans laquelle sont cherchées les réponses. Au travers de ces traitements, nous obtenons une structure informationnelle enrichie que nous pouvons interroger par un traitement minimal de la question. Dans cet article, nous exposons d'abord pourquoi nous avons choisi d'effectuer des traitements sur les textes. Ensuite, nous présentons le traitement appliqué aux documents pour récolter des données destinées à la construction de la structure informationnelle. Puis nous décrivons brièvement la technique d'interrogation de la structure. Enfin, nous présentons un évaluation de ce système de question-réponse.

2. Effectuer des traitements sur les textes plutôt que sur les questions

La décision de procéder à des traitement d'expansion sur les documents plutôt que sur la question provient d'une constatation : de nombreux enrichissements appliqués à la question ne correspondent pas à la signification de l'élément d'information qu'ils enrichissent dans le contexte qu'il occupe. Dans la figure 1, on voit que la méconnaissance du sens d'une unité lexicale conduit à effectuer une expansion erronnée dans deux cas (*universel* et *vague*) pour une expansion correcte dans un seul cas (*commandant*). Or si le sens correct de l'unité lexicale ainsi enrichie par synonymie était connu, l'expansion serait correcte et des erreurs pourraient être évitées.

Nous avons tiré les mêmes conclusions pour les autres types d'enrichissements : racinisation,

Question proposée :
Qui est le général des Perses ?

Synonymes de général :
commandant (sens n°2)
universel (sens n°1)
vague (sens n°3)

Question enrichie par tous les synonymes :

Qui est le	général commandant universel vague	des Perses ?

Question enrichie par les synonymes propres au sens :

Qui est le	général commandant	des Perses ?

Figure 1. Expansion synonymique : intérêt de la sémantique.

utilisation de dérivés, etc. Chacune de ces méthodes d'expansion rencontre le problème de la bonne compréhension de la signification des éléments d'information de la question.

Or lorsque l'expansion de la requête apporte des enrichissements qui ne correspondent pas à la signification des éléments qui constituent cette requête, ces enrichissements erronés sont susceptibles d'amener des réponses incorrectes. Cependant, l'utilisation d'un système capable de déterminer le sens exact de chaque élément composant la question - c'est-à-dire un système de désambiguïsation sémantique - ne semble pas capable de résoudre le problème. En effet, ces systèmes peuvent déterminer le sens des unités lexicales dans un énoncé par l'analyse du contexte, notamment lexical et syntaxique. Le contexte des questions est généralement court, et de plus les questions utilisent le plus souvent une syntaxe qui leur est particulière. Comme les documents permettent le plus souvent de disposer d'un contexte plus large que les questions, et que leur syntaxe est généralement plus conforme aux possibilités des grammaires des analyseurs syntaxiques, nous avons jugé qu'ils étaient plus propices à l'analyse sémantique, et donc à l'expansion d'énoncés.

Nous avons par ailleurs noté que les procédures appliquées aux questions étaient lourdes et contraignantes. De plus, au plus le système est subtil et les résultats de qualité, au plus les traitements appliqué aux questions sont substantiels. Mais l'importance de ces traitements est inversement proportionnelle à l'utilisabilité du système. Cependant, si les traitements les plus lourds sont appliqués aux documents avant l'utilisation du système plutôt qu'à la question lors de l'interrogation, la réponse à cette question peut être presque instantanée. Pour ces raisons, nous avons choisi d'appliquer les traitements aux documents de la base textuelle, d'enrichir ces documents et de conserver un index tant des données ajoutées que des informations originelles.

3. Traitements appliqués aux documents

Notre méthodologie est centrés sur des techniques linguistiques et sur l'utilisation de ressources lexicales. Nous considérons que l'information présente dans les textes est exprimée par les mots

« significatifs », qui sont les éléments d'information, par les relations syntaxiques qui relient ces éléments entre eux, et par le sens que ces mots possèdent dans le contexte où ils apparaissent, qui correspond à la signification des éléments d'information. De ce fait, l'analyse des textes consiste en une identification des unités lexicales qui les composent, des relations syntaxiques qui y apparaissent et du sens des mots.

Pour l'analyse morphologique et l'extraction des liens syntaxiques entre les unités lexicales, nous utilisons des outils conçus au *Xerox Research Centre Europe* (XRCE) de Grenoble. L'analyseur morphologique s'appelle NTM (*Normalizer - Tokenizer - Morphological analyser*) (Aït-Mokhtar, 1998 ; Trouilleux, 2001). Cet analyseur met en concurrence plusieurs transducteurs qui contiennent le lexique de la langue, des règles de génération qui constituent le « devin » (*guesser*), des règles de flexion morphologique, etc. Il permet d'obtenir la segmentation du texte en unités lexicales et propose pour chacune toutes les analyses morphologiques possibles. Par ailleurs, toute information ajoutée au transducteur lexical peut se retrouver comme une étiquette supplémentaire dans les analyses proposées.

L'analyseur syntaxique mis en œuvre est XIP (*Xerox Incremental Parser*) (Roux, 1999 ; Aït-Mokhtar *et al.*, 2002 ; Hagège et Roux, 2002). Cet analyseur robuste crée des syntagmes minimaux (*chunks*) et extrait les dépendances syntaxiques qui existent entre la tête des différents syntagmes minimaux. Pour effectuer ces opérations, il exploite des règles grammaticales. Les règles exploitent le contexte : catégorie grammaticale des unités lexicales, nature des syntagmes minimaux (nominaux, verbaux, prépositionnels, etc.) ou autres traits. Les grammaires XIP sont très souples, et cette flexibilité permet d'ajouter de nouvelles règles, de nouveaux types de dépendances, de nouveaux traits portant soit sur des mots, soit sur des dépendances ou encore sur des syntagmes minimaux, selon les besoins. En particulier, il est possible d'affecter des traits sémantiques sur des unités lexicales et d'en confier la gestion à une grammaire spécifique.

Le désambiguïsateur sémantique que nous utilisons est une évolution du système développé naguère à XRCE (Segond *et al.*, 2000 ; Brun *et al.*, 2001). Ce système exploitait les données extraite des exemples et des collocateurs du dictionnaire bilingue *Oxford Hachette French Dictionary* (Correard et Grundy, 1994) comme une information lexico-syntaxique de désambiguïsation. Le nouveau système de désambiguïsation sémantique (Jacquemin *et al.*, 2002 ; Jacquemin, 2003) fonctionne toujours sur base de règles de désambiguïsation obtenues à partir de l'information d'un dictionnaire, le *Dictionnaire des verbes et des mots français* (*Dubois*) (Dubois et Dubois-Charlier, 1997). Le système construit d'abord les règles de désambiguïsation sémantique, exploitant pour ce faire l'information lexicale et syntaxique contenue dans le dictionnaire (cf. exemple 2 page 637). Les règles de désambiguïsation sémantique sont créées selon le formalisme des grammaires XIP. Elles proviennent de schémas syntaxiques ou d'exemples qui sont attribués par sens aux entrées du dictionnaire.

La figure 2 présente une règle de désambiguïsation pour le sens n°2 « gagner » du verbe *remporter*. Cette règle permet de déterminer que dans un contexte où l'objet direct (`VARG[DIR]`) de *remporter* est *victoire*, le sens du verbe *remporter* est *gagner*, et pas *ramener ce que l'on avait apporté* (sens n°1).

Donc l'application de la désambiguïsation sémantique est effectuée par l'analyseur XIP et une grammaire composée des règles de désambiguïsation sémantique dès lors que les dépendances syntaxiques ont été extraites. L'application de ces règles de désambiguïsation assigne à chaque unité lexicale désambiguïsée un trait correspondant au numéro de sens qui lui est propre dans le contexte. De plus, les traits sémantiques (domaine d'application, classe sémantique) correspondant à ce sens dans le dictionnaire sont associés à cette unité lexicale.

Exemple extrait du Dubois :
On remporte la victoire sur ses adversaires (sens n°2 : « gagner »)

Une dépendance impliquant remporter :
`VARG [DIR](remporter,victoire)`

Règle conditionnelle de désambiguïsation :
```
remporter : VARG[DIR](remporter,victoire) ==>
sens « gagner »
```

Figure 2. Exemple d'extraction d'une règle de désambiguïsation sémantique.

Le résultat de l'analyse morpho-syntaxique et de la désambiguïsation sémantique sont stockés dans un index afin d'obtenir une structure informationnelle des textes traités. Chaque dépendance, chaque mot significatif, chaque trait est conservé dans l'index et permet d'aboutir aux fragments de textes où ils apparaissent dans la base documentaire.

4. Enrichissement de la structure informationnelle

En plus des traits sémantiques, c'est-à-dire le numéro de sens, la classe sémantique et le domaine d'application provenant du dictionnaire, qui ont été adjoints aux unités lexicales désambiguïsées dans la structure informationnelle, d'autres méthodes d'enrichissement ont été réalisées pour fournir autant d'actualisation que possible aux éléments d'information. Par cette méthode, les données présentes dans les questions pourront coïncider avec les mêmes données présentes sous une autre forme dans les documents.

La première méthode d'expansion des textes consiste à y introduire des synonymes qui proviennent d'un ou plusieurs dictionnaires de synonymes. Pour ce faire, nous utilisons les « parasynonymes » fournis par le *Dubois*, qui sont attachés au sens désigné par le désambiguïsateur sémantique. Mais ces parasynonymes ne sont pas assez nombreux : il y a au maximum deux synonymes par sens dans ce dictionnaire. Comme nous disposions d'autres dictionnaires, nous avons pu utiliser les synonymes qu'ils fournissaient pour enrichir les textes.

Ces dictionnaires sont d'une part un dictionnaire de synonymes (le *Dictionnaire des synonymes* de René *Bailly* (1947)), d'autre part des dictionnaires sémantiques (*Memodata* et *EuroWordNet* français). Le problème est que les synonymes ne sont pas distribués selon les sens du *Dubois* dans ces ressources, et que l'enrichissement ne peut donc se faire selon la signification des unités lexicales originelles déterminée par la désambiguïsation sémantique. Pour pouvoir conserver un enrichissement sémantique, seuls les synonymes qui possèdent les mêmes traits sémantiques dans le *Dubois* que le mot désambiguïsé peuvent être utilisés pour enrichir ce mot. Les autres synonymes proposés sont considérés comme des synonymes pour un autre sens.

La méthode d'enrichissement consiste simplement à mettre en disjonction, dans la structure de l'information que l'on est en train de construire, l'unité lexicale originelle et le ou les synonymes qui lui correspondent dans l'énoncé. La figure 3 permet de comprendre le fonctionnement de cet enrichissement. La désambiguïsation sémantique a déterminé que le sens de *chef* dans l'énoncé est « autorité » (n°1). Les synonymes proposés pour ce sens (en gras) viennent simplement se glisser en disjonction dans les dépendance où le mot originel apparaît.

L'autre manière d'enrichir les documents consiste à utiliser des formes dérivées des unités lexi-

Énoncé à enrichir :
César fixe à Alésia le **chef** des coalisés

Synonymes proposés :
commandant (sens 1 : autorité)
dirigeant (sens 1 : autorité)
cuisinier (sens 2 : responsable de cuisine)
maître queux (sens 2 : responsable de cuisine)

Dépendances impliquant chef (sens 1, structure non enrichie) :
`VARG[DIR](fixe,chef)`
`NMOD[INDIR](chef,de,coalisés)`

Dépendances impliquant chef (sens 1, structure enrichie) :

$$\texttt{VARG[DIR]}\left(\texttt{fixe},\ \begin{array}{c}\texttt{chef}\\ OU\\ \textbf{commandant}\\ OU\\ \textbf{dirigeant}\end{array}\right)$$

$$\texttt{NMOD[INDIR]}\left(\begin{array}{c}\texttt{chef}\\ OU\\ \textbf{commandant}\\ OU\\ \textbf{dirigeant}\end{array},\ \texttt{des},\ \texttt{coalisés}\right)$$

Figure 3. Enrichissement synonymique.

cales qui composent les énoncés, considérant que la forme dérivée a un sens proche de la forme originelle. Le *Dubois* possède un champ informationnel qui indique quels sont les dérivés d'une unité lexicale dans chacun de ses sens. Par exemple, le dérivé *tracteur* provient du mot *tirer*. Mais si *tracteur* dérive bien de *tirer* sous son sens « remorquer », il n'en va pas de même sous son sens « faire partir une arme à feu ».

Dans la plupart des cas, le dérivé appartient à une catégorie grammaticale différente de celle du mot originel. De ce fait, les relations syntaxiques que le mot originel entretient avec son environnement ne peuvent rester inchangées lors de la dérivation.

En examinant chaque type de dérivation suffixale, nous avons construit une table de correspondance indiquant le schéma syntaxique correspondant à chaque contexte originel. Ce tableau de correspondance permet de transformer un fragment de texte par dérivation d'une unité lexicale sans en modifier notablement le sens. Par exemple, l'énoncé l'*arrivée du train* peut être transformé par un dérivé *arriver* grâce au schéma syntaxique correspondant (`NMOD[INDIR](arrivée, de, train) ==> SUBJ(arriver,train)`). L'énoncé qui en découle *le train arrive* a pratiquement la même signification que celui dont il découle.

5. Interroger la structure informationnelle

Même si la base documentaire a été analysée et qu'une expansion l'a considérablement enrichie, la phase d'analyse de la requête seule permet d'identifier les données proches de la réponse à la

question. Le traitement de la requête est donc de première importance. Ce traitement est similaire à celui des documents, mais il est limité à l'analyse linguistique, et encore cette dernière ne comprend-elle pas la désambiguïsation sémantique. La procédure d'expansion n'est pas nécessaire car la structure informationnelle a déjà été enrichie, et de plus elle ne peut être mise en œuvre par manque de contexte permettant l'application de la désambiguïsation sémantique.

L'analyse linguistique de la question amène à la construction d'une structure informationnelle locale, similaire à celle des documents, mais évidemment plus rudimentaire. La grammaire employée pour la question a été modifiée de telle manière que les caractéristiques spécifiques à l'interrogation n'apparaissent pas dans l'analyse, et donc pas non plus dans la structure locale. Si l'interrogation porte sur un mot explicite de la question, ce mot (appelé *focus*) disparaît de la structure locale, mais ses caractéristiques sémantiques sont maintenues sous forme de traits, et son requises dans la réponse. Dans les autres cas, les traits sémantiques du *focus* dépendent de l'interrogatif.

Une fois la structure locale élaborée, la réponse peut être trouvée dans la structure informationnelle documentaire par simple comparaison avec cette structure locale. Le système peut être interrogé à différents niveaux (texte, paragraphe, phrase), mais dans le cas d'une application de question-réponse, seul le niveau de la phrase est pertinent. Une réponse congrue est une phrase qui contient les informations contenues dans la structure locale, ainsi que l'élément non identifié qui correspond aux traits sémantiques du *focus*.

Question proposée :
De quel chef Domitien fut-il le successeur ?

Structure de la question (dépendances) :
NMOD[SPRED](**Domitien,successeur**)
NMOD[INDIR](successeur,de,chef)

Réponse attendue :
(...) Domitien succéda à l'empereur Titus (...)

Structure enrichie du texte :
SUBJ(succéda *OU remplacer*,Domitien)
VARG[INDIR](succéda,à,empereur)
VARG[DIR](remplacer,empereur OU *chef)*
NN(empereur OU *chef OU souverain*,Titus)
NMOD[INDIR](successeur,de,empereur *OU* ***chef)***
NMOD(Domitien,successeur)

Figure 4. Interrogation d'un texte enrichi.

La figure 4 montre comment l'expansion du texte permet de mettre en correspondance une question avec la réponse dans laquelle la plupart des termes de la requête ne sont pas représentés. Les dépendances et leurs arguments forment la structure informationnelle. Les parties en gras montrent les éléments qui se répondent dans la structure de la question et dans celle du texte. Les éléments en italiques sont issus de l'enrichissement, soit synonymique, soit dérivationnel.

6. Évaluation

Étant donné qu'il n'existe aucune évaluation standard de la tâche de question-réponse pour le français, nous nous sommes inspiré du modèle de TREC-8 (Voorhees, 1999). Pour notre évaluation, huit personnes non spécialistes on lu 50 articles de l'*Encyclopédie Hachette Multimédia 2000* et on proposé 25 questions chacun sur leur lecture. Les 200 questions sont en français correct, sans erreur orthographique ni syntaxique. Chaque question possède au moins une réponse dans les articles. Nous avons établi un « plancher » (*baseline*) en effectuant une interrogation de la base documentaire en utilisant uniquement le vocabulaire (lemmatisé) présent dans les documents, sans autre traitement. Nous établissons un second point de comparaison en enrichissant les documents avec tous les synonymes sans aucune utilisation de l'analyse linguistique, et sans filtrage par la désambiguïsation sémantique (*synonymes sans sémantique*).

Nous comparons les résultats avec l'enrichissement sémantique synonymique (*synonymes avec sémantique*), puis avec tous les enrichissements (*tous les enrichissements*). Seules les cinq premières réponses sont prises en compte. Une score est alloué à chaque question. Ce score est de 1 si la première réponse est correcte, de 1/2 si la première réponse est incorrecte et la deuxième est correcte, de 1/3 si les deux premières réponses sont fausses et que la troisième est correcte, etc. Si aucune bonne réponse n'apparaît parmi les cinq premières, le score est de 0. Les scores indiqués dans le tableau d'évaluation sont une moyenne pour les 200 questions proposées.

Enrichissement	Score	Pas de réponse
Plancher	0.295	139
Synonymes (sans sémantique)	0.303	137
Synonymes (avec sémantique)	0.487	100
Tous les enrichissements	0.504	97

Cette évaluation montre la réelle amélioration de l'utilisation d'une discrimination sémantique pour l'expansion d'énoncés appliquée à la tâche de question-réponse. Ces enrichissements permettent d'éliminer de fausses propositions de réponses et d'obtenir plus rapidement la réponse juste à une question posée.

Malgré ces résultats prometteurs, de nombreuses améliorations peuvent encore être effectuées pour améliorer les performances du système. Notamment, le manque de synonymes est encore flagrant lorsqu'on étudie les questions et les (manques de) réponses en profondeur. Par exemple, *bru* et *belle-fille* ne sont jamais mis en relation. D'autres relations sémantiques devraient également être intégrées pour permettre de faire le lien entre des informations proches, mais pas synonymes : l'hypéronymie permettrait ainsi de mettre en relation *fonction* et *consul*, le premier recouvrant le second. Il y a encore la nécessité de marquer l'équivalence entre certaines expressions, comme *X est fatal à Y* qui équivaut à *Y meurt de X*.

Par ailleurs, l'exigence d'une correspondance stricte et exacte entre la structure de la question et celle des documents elimine de fait des réponses légèrement différentes de la structure exigée. Une méthode de dégradation de la correspondance des structures, avec pondération des réponses candidates, serait sans doute la bienvenue.

7. Conclusion

Dans cet article, nous avons montré l'intérêt du prétraitement des documents contenus dans une base documentaire destinée à être interrogée selon les modalités de la tâche de question-réponse. Le fait de soumettre le texte à des traitements plutôt que la question est justifié par le contexte

généralement plus large dans les énoncés des documents, rendant de ce fait possible l'usage d'un système de désambiguïsation sémantique pour identifier le sens correct de chaque élément d'information dans le contexte dans lequel il apparaît. De ce fait, les différents enrichissements apportés pour réaliser une expansion de texte sont plus précis et les réponses apportées sont plus souvent correctes.

Nous présentons également un système de question-réponse fondé sur la construction d'une structure informationnelle d'une base documentaire. Cette structure provient de l'analyse linguistique (morphologique, syntaxique et sémantique) des documents, et de l'expansion des énoncés par enrichissement synonymique et dérivationnel. L'évaluation confirme l'intérêt de l'enrichissement sémantique, qui est rendu possible seulement par le traitement des textes plutôt que celui des questions.

Références

Aït-Mokhtar S. (1998). *L'analyse présyntaxique en une seule étape*. Thèse de doctorat, Université Clermont 2 Blaise Pascal.

Aït-Mokhtar S., Chanod J.-P. et Roux Cl. (2002). Robustness beyond shallowness : incremental deep parsing. *Natural Language Engineering*, vol. (8/2-3) : 121-144.

Bailly R. (1947), *Dictionnaire des synonymes de la langue française*. Larousse.

Brun C., Jacquemin B. et Segond Fr. (2001). Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale. *Traitement Automatique des Langues*, vol. (42/3) : 667-690.

Corréard M.-H.et Grundy V. (Éds) (1994). *The Oxford-Hachette French Dictionary*. Oxford University Press.

Dubois J. et Dubois-Charlier Fr. (1997). *Dictionnaire des verbes français*. Larousse. Ce dictionnaire n'existe au départ qu'en version électronique. Il dispose de son complément naturel *Dictionnaire des mots*. Ces deux ressources sont disponibles à ELDA.

Ferret O., Grau Br., Hurault-Plantet M., Illouz G. et Jacquemin Ch. (2002). Quand la réponse se trouve dans un grand corpus. *Revue d'Ingénierie des Systèmes d'Information*, vol. (7/1-2) : 95-123.

Hagège C. et Roux Cl. (2002). A Robust and Flexible Platform for Dependency Extraction. In *Proceedings of LREC 2002* : 520-523.

Harabagiu S., Moldovan D., Psca M., Mihalcea R., Surdeanu M., Bunescu R., Gîrju R., Rus V. et Morarescu P. (2002). *Falcon : Boosting Knowledge for Answer Engines*.

Hull D.A. (1999). Xerox TREC-8 Question Answering Track Report. In Voorhees E.M. et Harman D. (Éds), *Proceedings of The Eighth Text Retrieval Conference (TREC-8)* : 743-752.

Jacquemin B. (2003). *Construction et interrogation de la structure informationnelle d'une base documentaire en français*. Thèse de doctorat, Université Paris III Sorbonne Nouvelle.

Jacquemin B., Brun C. et Roux Cl. (2002). Enriching a text by semantic disambiguation for information extraction. In de Loupy Cl. (Éd.), *LREC 2002 Workshop Proceedings. Using Semantics for Information Retrieval and Filtering*.

Roux Cl. (1999). Phrase-driven parser. In *Proceedings of VEXTAL'99* : 235-240.

Segond F., Aimelet E., Lux V. et Jean C. (2000). Dictionary-driven Semantic Look-up. In *Computer and the Humanities. Special Issue on SENSEVAL*, vol. (34/1-2) : 193-197.

Trouilleux Fr. (2001). *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. Thèse de doctorat, Université Clermont 2 Blaise Pascal.

Voorhees E.M. (1999). The TREC-8 Question Answering Track Report. In Voorhees E.M. et Harman D. (Éds), *Proceeding of the Eighth Text REtrieval Conference (TREC-8)* : 77-82.

Voorhees E.M. et Harman D. (Éds) (1999). *Proceeding of the Eighth Text REtrieval Conference (TREC-8)*.

Paradoxes, dilemmes et contradictions : une mise en lumière au moyen du logiciel EVOQ©

Jean-Marie Jacques, Nathanaël Laurent, Anne Wallemacq

Facultés Universitaires Notre-Dame de la Paix ReCCCoM (Namur) – Belgique
nathanael.laurent@fundp.ac.be

Abstract

This paper focuses on the analysis of language in the realm of organizational settings. We present an innovative software (EVOQ©) and the theoretical frameworks on which it is based. we also explore ways to make this software operational in the field of organization studies. An examples is used to show how EVOQ© is able to highlight organizational paradoxes which affect the semantic fields woven (enacted) in discourse.

Résumé

Ce papier porte sur l'analyse du discours dans les organisations. Nous présentons un logiciel novateur (EVOQ©) les cadres théoriques qui le fondent. Nous explorons une manière de rendre opérationnel ce logiciel dans le champ de l'analyse des organisations. Un exemple montre de quelle manière EVOQ© est capable de mettre en lumière des paradoxes organisationnels qui affectent les champs sémantiques tissés (énactés) dans le discours.

Mots-clés : analyse du discours, organisations, logiciel, paradoxes, champ sémantique.

1. Le langage et l'organisation

Notre logiciel EVOQ© se situe dans le champ scientifique de l'analyse des organisations, autour de la question du langage (Grant *et al.*, 1998). Quatre courants peuvent y être dégagés, il existe entre eux croisements et recouvrements (Westswood et Linstead, 2001).

Le premier de perspective Foucauldienne s'intéresse à l'articulation entre pouvoir et discours, montrant combien la production d'un champ discursif et d'un champ de connaissance coïncide avec un processus d'assujettissement (Hassard 1993 ; Linstead, 2001). Par exemple, le développement d'un champ de discours autour de la « gestion des ressources humaines », à la fois constitue un « sujet » de recherche et d'action : « les ressources humaines » et les outils de sa domination (Townley, 1994).

Le deuxième d'inspiration anthropologique, s'intéresse à la production langagière des organisations en tant que produit « culturel » : en quoi le langage créé dans cette organisation réorganise la codification du réel et plus loin, réorganise la perception des membres de celle-ci. L'approche est ici constructiviste (Berger et Luckmann, 1986 ; Whorf, 1956) et phénoménologique : comment une organisation, et les individus qui la composent organisent leur perception du réel et le saisissent comme univers de sens (Weick, 1979, 1992).

Le troisième, proche du précédent dans ses présupposés théoriques s'intéresse aux récits dans l'organisation. Comment l'organisation donne sens à ce qu'elle est et construit son histoire en se racontant et se re-racontant (Czarniawska, 1998). Plus loin, l'organisation elle-même, plutôt que d'être vue comme un cadre préexistant, est vue comme un discours qui lui donne son

unité, sa vie et sa continuité. L'organisation n'est plus vue ici comme siège de discours, comme étant une production discursive.

Enfin un courant « managerial » met à plat des manières dont les membres d'une organisation raisonnent. On trace ici des « cartes cognitives », individuelles ou collectives, qui permettent aux décideurs de réfléchir sur leurs propres présupposés. L'accent est mis sur le raisonnement « ordinaire », ses boucles, ses simplifications possibles (Eden, 1992 ; Wallemacq et Jacques, 2003).

2. Le logiciel *Evoq*©

Le logiciel EVOQ© reprend au dernier courant l'idée de *carte cognitive*. Il met en effet à plat les termes de référence individuels ou collectifs dans lesquels raisonnent les individus d'une organisation. Ceci étant, l'accent est mis moins sur le raisonnement proprement dit que sur la perception, la manière dont le langage organise la vision du monde propre à l'organisation et en quoi, bien souvent, cette vision du monde est « enactée »[1] (Varela, 1993 et 1996). En cela, elle s'inscrit dans le deuxième courant. Enfin, la vocation du logiciel peut être clairement critique et peut donc s'articuler sur le premier courant : comment des termes de référence oblitèrent d'autres termes de références possibles et comment une vision du monde est « naturalisée » ? Plus proche de l'action, en mettant à plat ces termes de référence, EVOQ© invite à une prise de conscience de ceux-ci et à une analyse critique de leurs possibilités et contraintes. Ainsi nous avons beaucoup travaillé sur les termes de référence qui prévalent dans la gestion des crises (Jacques *et al.*, 2002 ; Wallemacq et Jacques, 2003).

Ceci étant, sur un plan différent mais lié, le logiciel est aussi pour nous l'occasion de réfléchir plus avant sur la logique des langages naturels. Fidèle à la tradition post Wittgensteinienne qui ne fait plus du langage naturel un genre de dérivé d'un langage formel, nous nous attachons à étudier la conformation propre du langage naturel, tel qu'il est produit et utilisé dans les organisations. Le logiciel met l'accent moins sur le raisonnement que sur la « perception ». Il part d'une conception inspirée par les structuralistes (Greimas, 1966 ; Hiernaux, 1977) au départ, mais surtout par sa relecture phénoménologique (Merleau-Ponty, 1960 ; Wallemacq et Jacques, 2003). Loin d'être un moyen de communication d'un locuteur conscient du message qu'il veut exprimer, le langage est vu comme un milieu ambiant, dans lequel le locuteur baigne. Il est cette matière langagière qui le porte, le submerge, le dépasse et avec lequel il compose au quotidien, créant parfois des effets de sens non délibérément recherchés par lui. Ce milieu est composé de la chaîne des évocations qui gravitent autour des mots. Ces évocations ne sont ni vraies ni fausses, simplement elles sont inexorablement appelées dans un discours donné et constituent le code, le cadre de référence à l'intérieur duquel un discours sera reçu et prononcé.

Ce qui nous intéresse est de saisir les *principes d'articulation des évocations*. Tout laisse en effet à penser qu'elles ne s'encastrent pas dans le jeu vertical des boîtes qui s'imbriquent les unes dans les autres. Ainsi si dans la logique du dénoté, « poulet » fait partie de la catégorie « animal », dans la logique du connoté, des évocations, poulet et animal appellent des *champs sémantiques* différents (poulet = chair de poule, agent de police, élevage intensif, surnom affectueux, etc. ; animal = primaire, brutal, non évolué, sale, ou encore sauvage, non artificiel, etc.). Les aspects fonctionnel d' EVOQ© seront abordés dans l'exemple *infra*.

[1] « […] la cognition […] est fondée sur l'activité concrète de tout l'organisme […]. Le monde n'est pas quelque chose qui nous est donné : c'est une chose à laquelle nous prenons part en fonction de notre manière de bouger, de toucher, de respirer et de manger. […] l'énaction connote cette production par la manipulation concrète. » (Varela, 1996 : 23-24).

3. Mise en lumière des paradoxes : un apport opérationnel du logiciel *Evoq*©

Dans cette approche, un des points qui nous est apparu plus clairement, et que nous trouvons intéressant d'exploiter, est que le logiciel met régulièrement à jour les contradictions, ou les paradoxes dans lesquels une organisation et les individus qui la composent se débattent. Ceci fournit un éclairage nouveau sur la vie des organisations. Nous sommes à présent convaincus que les paradoxes et contradictions que l'on trouve dans la production discursive d'un locuteur donné (au sein d'une organisation) n'est bien souvent pas rapportable à un « défaut de logique » dans le chef de celui-ci, mais est le résultat systématique d'une production organisationnelle. Les organisations se débattent dans des paradoxes qui sont relayées par leurs membres. On peut parfois même pousser le propos plus loin en disant que certaines organisations sont bâties sur un discours paradoxal. Une bonne partie de l'énergie des membres consiste à balancer dans les termes de la contradiction, à tourner dans le paradoxe ou à lui trouver des solutions tant sémantiques que pratiques. La créativité organisationnelle est ici assez surprenante.

Quand nous parlons d'« organisation », nous ne postulons pas une entité préalable et réifiée. Nous pointons simplement qu'il y a un niveau collectif qui ne résulte pas de la volonté ou d'un accord conscient et délibéré entre les acteurs. C'est, pour reprendre l'expression des ethnométhodologues, le caractère réflexif du code (Heritage, 1984). Celui-ci n'est jamais explicite mais il fait l'objet de présuppositions qui sont réciproques. Nous sommes assez proches de leur thèse disant qu'il n'y a pas d'abord des « membres » qui partagent le même langage, mais que c'est le partage du langage naturel (du code) qui fonde le « membership », l'appartenance à l'organisation vue comme une communauté langagière, comme parlant la même langue, comme présupposant que l'autre présuppose de la même manière ce qu'un énoncé « veut dire » au delà de sa compréhension littérale.

L'exemple *infra* montre combien Evoq© permet de mettre à jour des « paradoxes organisationnels ». La réflexion que nous proposons ici est de voir comment ses principes mêmes de construction et d'opération permettent de mettre à jour de tels paradoxes. Notre hypothèse en effet n'est pas que les organisations que nous avons analysées sont plus que les autres sujettes à de tels paradoxes mais que l'outil que nous utilisons permet de les mettre en lumière sans doute davantage qu'un autre. Nous tenterons d'analyser les différentes formes que revêtent les paradoxes débusqués et d'éclairer la logique organisationnelle qui d'après nous les produit dans la bouche des locuteurs.

Mais revenons tout d'abord rapidement à la définition que donne Watlzawick du paradoxe et qui, bien qu'ancienne, a le mérite de ne pas se cantonner aux seuls plans syntaxiques et sémantiques, mais ouvre la possibilité de « paradoxes pragmatiques ». Watzlawick reprend le concept de « double contrainte » et de communication paradoxale, montrant que celle-ci est un phénomène systémique. Il ne saurait dès lors pas se réduire au simple plan individuel (la volonté ou l' inaptitude du locuteur). Ce concept est évidemment utile dès que l'on veut analyser des paradoxes visiblement construits par un cadre organisationnel.

Watzlawick définit le paradoxe comme « une contradiction qui vient au terme d'une déduction correcte à partir de prémisses « consistantes » (Watzlawick *et al.*, 1972). Il distingue trois groupes de paradoxes :

– *Les paradoxes logico-mathématiques* qui correspondent au domaine de la *syntaxe logique*.
– *Les définitions paradoxales* qui relèvent du domaine de la *sémantique.*

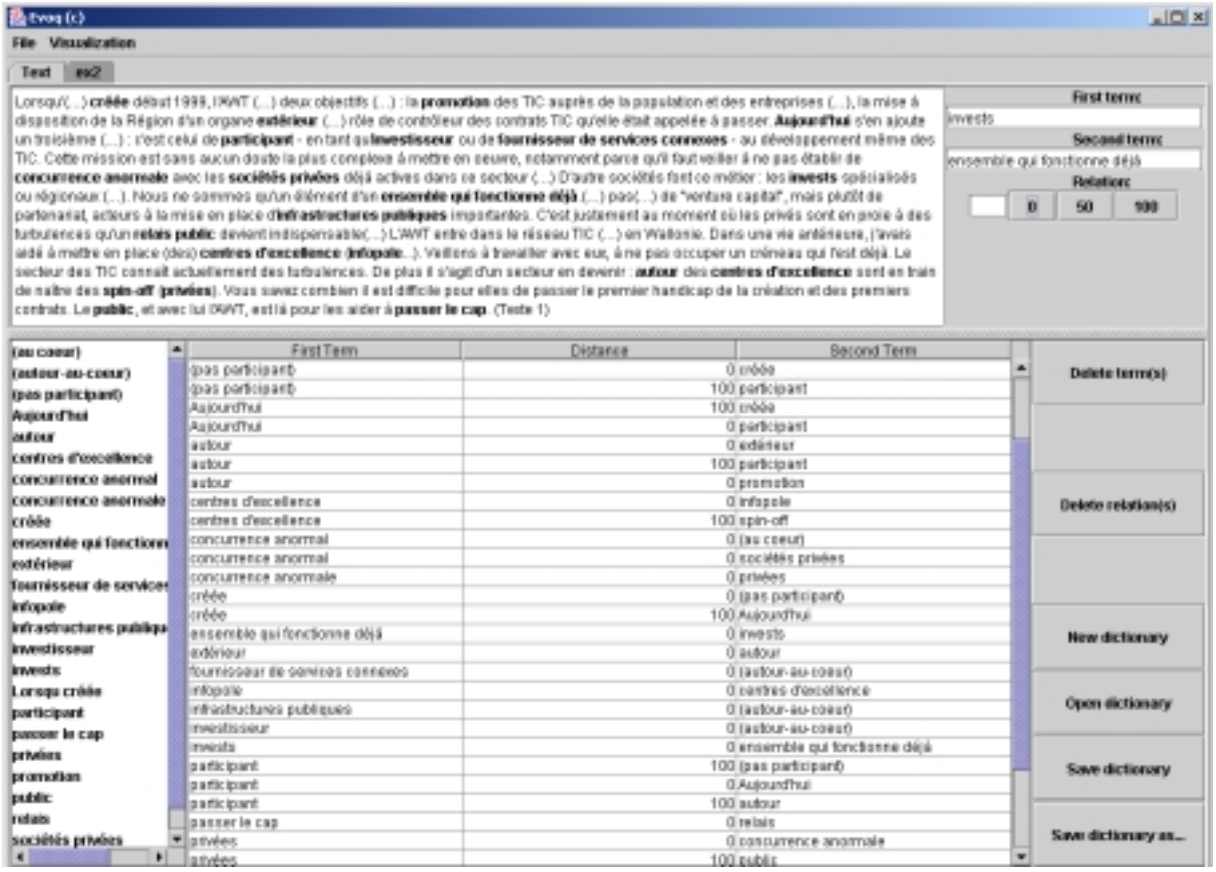
– *Les paradoxes pragmatiques :* Paradoxes du type « soyez spontanés » : il faudrait être spontané par obéissance qui est l'opposé de la spontanéité. Pour obéir il faut désobéir et réciproquement. D'un point de vue formel : « un message est émis qui est structuré de manière telle a) qu'il affirme quelque chose, b) il affirme quelque chose sur sa propre affirmation, c) ces deux affirmations s'excluent. Ainsi, si le message est une injonction, il faut lui désobéir pour lui obéir (Watzlawick *et al.*, 1972) ». Ces paradoxes aboutissent à une double contrainte lorsque ce message est prononcé dans le cadre d'une relation interpersonnelle importante et que le récepteur se trouve dans l'impossibilité de sortir du cadre et d'en démontrer l'impossibilité.

Le paradoxe se distingue de la contradiction (simple coexistence de deux termes contradictoires) ou de deux injonctions contradictoires (faites ceci et ne faites pas ceci). Le « ou » exclusif est remplacé par un « et ». Le paradoxe suppose que l'un des termes (blanc) implique par déduction qu'il est son opposé (non blanc), ou que l'obéissance à un des ordres (faites ceci) suppose par son exécution même sa négation (ne faites pas ceci). Le « ou » exclusif est remplacé par une implication. Dans la contradiction il reste la possibilité de faire l'un sans faire l'autre. Dans le paradoxe, le fait même de faire l'un me conduit à faire l'autre (qui est son opposé). Nous pouvons présenter un exemple de paradoxe organisationnel levé grâce à EVOQ©, il porte sur 2 textes issus d'une brochure de l'Agence Wallonne des Télécommunications (AWT)

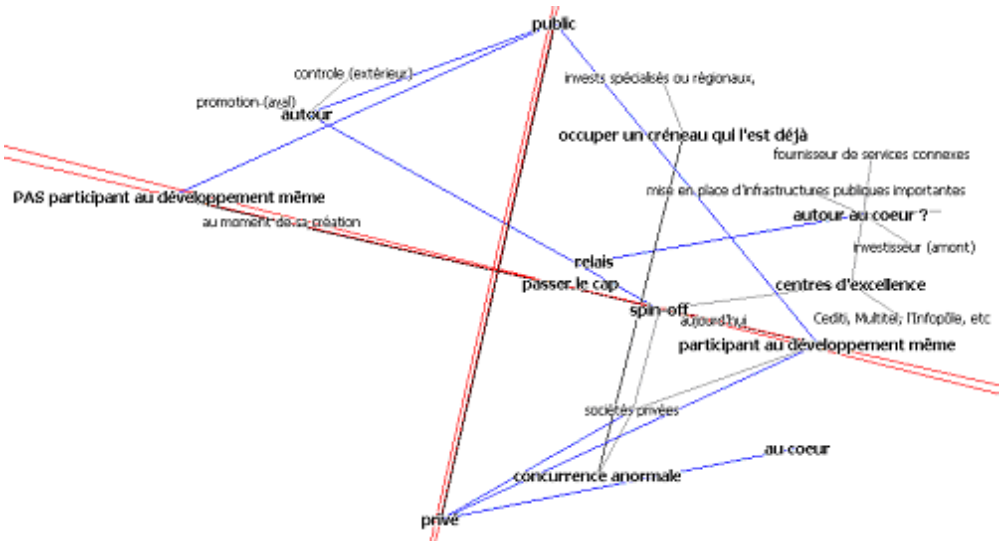
3.2. *Comment être autour et au centre à la fois*

LE CONTEXTE : l'AWT décrit dans une brochure ses objectifs.

LE LOGICIEL : reprend les disjonctions

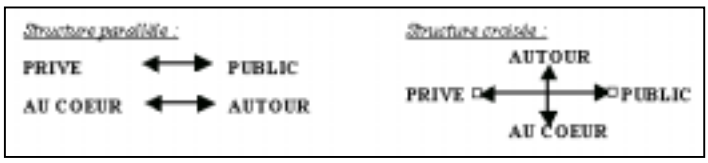


Le graph et l'analyse



Il s'agit de promouvoir les technologies wallonnes soit en amont (« organe extérieur de contrôle des contrats TIC ») soit en aval («promotion des TIC »). Le troisième objectif (participer au développement même de ces technologies) amène des difficultés. En effet, sur la première disjonction (être en aval, « au cœur même » / être en amont « autour ») vient se « coller » une seconde disjonction : le secteur public est « autour », le développement même étant laissé au secteur privé. S'en suit un problème de positionnement : comment être autour et au cœur en même temps. Car si on se positionne simplement « au cœur », on fait une concurrence anormale au secteur privé, ce qui n'est évidemment pas le but ; si l'on se trouve simplement « autour », bien d'autres instances publiques s'en chargent déjà. Il faut donc résoudre la quadrature du cercle et se résoudre à être « autour et au cœur à la fois ».

Des solutions sont utilisées : 1) Développer les centres d'excellence autour desquelles des spin-off vont naître (favoriser l'émergence [autour] d'une structure privée [au cœur]). 2) Mise en place d'infrastructures publiques : relais publics [autour] quand le secteur connaît des « turbulences » ; relais quand le centre n'est plus capable. 3) Aider les spin-off naissantes « à passer le cap des premiers contrats ».



Un système de disjonctions stables et parallèles prévaut dans un certain champ sémantique. Ainsi la disjonction privé/public se superpose assez classiquement avec le développement même de l'activité économique qui est laissée au secteur privé, le secteur public intervenant « autour » en support à cette activité. Pour l'AWT, la situation devient difficile quand, en tant qu'organisme public elle doit néanmoins intervenir dans le développement même de l'activité économique. C'est ici que le paradoxe se noue. Elle ne peut pas être « seulement autour », sinon elle ne remplit pas ses objectifs, d'une part, et elle rentre en concurrence avec d'autres organismes publics qui assurent déjà cette mission, d'autre part. Elle ne peut pas non plus être uniquement au cœur de l'activité économique car elle se substitue alors au privé et provoque avec eux une concurrence anormale. Elle est donc condamnée à ne pas pouvoir trancher entre « autour » et « au cœur ». Elle doit être « autour-aucoeur » ce qui est une position a priori intenable.

Le paradoxe est créé ici par l'ouverture d'une structure croisée là où la structure était parallèle, ouvrant une case « impossible ». Il s'agirait d'une simple contradiction au sens donné

plus haut (un « et » logique en lieu et place d'un « ou » logique) si ce n'est qu'ici il y a une pression forte pour tenir ensemble la contradiction car les deux autres cases (public/autour et privé/au cœur) sont déjà investies. L'on ne peut pas « choisir » entre les deux solutions. Il faut être « autour-aucoeur ».

Surgit dans un autre texte la « trouvaille » (tout à la fois sémantique et pratique) permettant de « résoudre » le paradoxe, d'être à la fois « autour et au cœur » : « l'incubateur numérique ». Si nous cherchons (sur Wordnet par ex.) les associées de incubateur on vois : milieu chaud, protégé, à l'intérieur, ensemble (l'extérieur par effet d'écho est froid, exposé, composé d'entités isolées). Ces associés et opposés « collent » déjà assez bien avec ceux du texte précédent dont les solutions étaient autant de tentatives pour résoudre le paradoxe. Le secteur public intervient en effet quand le contexte est menaçant, ou pour des investissements que les entreprises isolées ne peuvent faire, ou pour protéger les entreprises naissantes. Mais bien plus l'incubateur, au sens premier, est le lieu du développement par le seul effet du regroupement à l'intérieur, à la chaleur et en milieu protégé de petits éléments. Créer les conditions d'un développement spontané, quelle plus belle manière d'être à la fois autour et au cœur ? On emprunte ici les propriété sémantiques de l'incubateur, (l'ensemble des associés et opposés qu'il charrie) au profit des technologies wallonnes. On crée donc un espace protégé dans lequel les PME wallonnes pourront se regrouper créant par le fait même une génération spontanée (d'idées ? de produits ? de nouvelles petites entreprises sans doute). On voit ici l'importance de l'effet sémantique du mot incubateur.

Texte 2: Le ministre(...) peaufine son incubateur numérique wallon (...). Inspiré d'expériences similaires au Japon, il devrait mettre à la disposition des entreprises du secteur des nouvelles technologies numérique (...) des infrastructures coûteuses qu'aucune d'elles ne pourrait s'offrir seule. A la fois hôtel pour start-up numériques, vitrine technologique et centre d'archivage, l'incubateur accueillera des entreprises très diverses. Des start-up s'y mêleront avec des entreprises déjà en activité (...) vitrine technologique(...). Les entreprises spécialisées dans le multimédia, l'image et la vidéo numérique, la reconnaissance vocale, la télévision interactive et les technologies de protection des droits d'auteurs devraient trouver refuge dans ce nouvel incubateur. Quant au centre d'archivage, l'un des volets les plus novateurs du projet, il permettra d'étudier différentes technologies de conservation et de restauration d'archives, ainsi que la constitution de bases de données de documents numériques.(...) L'un des principaux enjeux politiques concerne évidemment l'implantation de l'incubateur, avec en toile de fond les perspectives d'emploi (...) A ce stade, le budget est de 17,5 millions d'euros (...). Rien n'indique cependant que ce budget public ne sera pas complété par des fonds privés. Plusieurs opérateurs industriels et financiers (dans le secteur des télécoms, notamment), auraient déjà émis le souhait d'être présents sur le site de l'incubateur, que ce soit sous la forme d'apports en nature ou en capital. Le projet, préparé par l'Agence wallonne des télécoms, a été discuté jeudi(...).

WordNet 2.0 Search

Results for "Derivationally related forms" search of noun "incubator"

Display derived forms -
nouns and verbs that are related morphologically.
Each related synset is preceeded by its part of speech. Each word in the
synset is followed by its sense number.

```
1 sense of incubator


Sense 1
incubator, brooder --
 (apparatus consisting of a box designed to
 maintain a constant temperature by the use
 of a thermostat; used for chicks or premat
ure infants)
       RELATED TO->(verb) incubate#2
           => brood, hatch, cover, incubate
 --
 (sit on (eggs); "Birds brood"; "The female
 covers the eggs")
```

4. Conclusions

Avec EVOQ©, notre approche n'est pas simplement méthodologique mais repose toute aussi sur des principes épistémologiques et mêmes ontologiques (Wallemacq et Jacques, 2004). Ces principes- impliquent de considérer le langage comme consistant en soi et ne renvoyant pas à un « ailleurs » (une réalité en soi, la volonté expressive d'un acteur, une organisation) dont il serait la simple expression. La perspective d'EVOQ© consiste à considérer que le langage n'est pas second ou dérivé, mais qu'il participe activement à la constitution de ce que nous appelons réalité, organisation ou individu.

Dans cette conclusion, nous voulons revenir sur l'implication des paradoxes, lorsqu'on quitte le simple registre langagier. Qu'implique la présence de ces paradoxes pour l'expérience subjective des individus d'une part, que nous enseignent-ils sur la vie des organisations d'autre part ? Nous sommes immanquablement conduits à ces questions par la nature « pragmatique » ou « organisationnelle » des paradoxes que nous avons présentés.

Une entreprise, un laboratoire, une confrérie, une famille, une communauté, etc. sont autant de groupes possédant un langage commun et des filières d'apprentissage de ce langage. Au delà de ce langage, y sont également associé les expériences subjectives qui l' accompagne. Cette approche s'effectue au travers des éléments sémantiques contenus dans le discours individuel et/ou collectif. A l'aide du logiciel EVOQ©, le chercheur se penche sur la « langue naturelle ». Que se passe-t-il si nous admettons que celle-ci ouvre sur l'expérience subjective ?

Au plan de l'analyse, un triple mouvement d'inférence se met en place :

– *Du discours « collectif » vers l'expérience subjective « individuelle »*: le chercheur part de ce que la personne emploie comme formes ou jeux de langage ou comme (cf. Wittgenstein) par lesquels il peut tendre vers l'expérience subjective et intersubjective de l'individu.
– *Du discours « individuel » vers l'expérience subjective « collective »* : Mais inversement, accéder à la compréhension d'un groupe (à son langage commun, à ses expériences paradigmatiques associées), c'est accéder aux filières d'apprentissage qui font adopter son langage, langage qui permet à ses membres d'utiliser les mêmes exemples et expériences paradigmatiques comme autant de sources nouvelles d'expériences subjectives associées.
– *De l'expérience subjective collective vers l'expérience subjective individuelle*. Au cours du discours intersubjectif (interview), la personne utilise le discours collectif. Elle reproduit alors un message partiellement connecté à sa propre expérience subjective, c'est-à-dire partiellement « percolé » qui, sans forcément que la personne en soit consciente, véhicule un jeux de langage formé par l'expérience collectivement vécue.

C'est là que, avec l'aide d'EVOQ© il peut s'avérer très utile d'aller au delà des mots, et plus précisément « entre les mots ». Les niveaux structurel, relationnel et culturel, dans lesquels baignent toute organisation, font passer de l'objectif au subjectif, du théorique au pratique, et ce en tirant parti du langage commun propre au groupe au sein duquel l'apprentissage s'opère. Evoq© permet donc d'aller au delà d'une simple analyse structurelle, mais permet de *pénétrer le sens* (donc ce qui résulte, émane, « percole » des expériences subjectives). L'objectif est de représenter une partie du monde propre au locuteur et au sein duquel le processus de création de sens s'est produit (Latiers *et al.*, 2004 ; Wallemacq et Jacques, 2003). Il s'agit là d'un effort de reconstruction explorant les filières de création du discours, du sens que le locuteur énacte de son vécu. On atteint ainsi à la compréhension de la personne en tant qu'elle évolue dans un niveau culturel qui englobe les niveaux structurels et relationnels

Voici une réflexion, en guise de conclusion, sur la manière dont ils se composent dans l'exemple que nous avons analysés :

– Les niveaux collectif et individuel ne sont pas des instances que l'on pourrait séparer (l'organisation d'une part, l'individu d'autre part) : on a vraiment l'impression que le paradoxe n'est pas un simple problème logique que les individus auraient à résoudre. Le champ sémantique dans lequel il se meuvent (et qui est collectif) est en lui-même paradoxal. C'est un champ sémantique qui n'est pas en dehors d'eux, mais le milieu dans lequel ils baignent, dans lequel il tournent et dans lequel ils se débattent,
– A la différence des paradoxes pragmatiques qu'isole Watzlawick, le paradoxe analysé ici ne résultant pas d'une « communication paradoxale ». Ici (la chose mériterait sans doute

d'être davantage élaborée) le paradoxe ne se noue pas au niveau du système d'interaction. Tout laisse à penser qu'il se noue à un « niveau » plus fondamental : la nature même de l'organisation est impliquée qui sont en soi contradictoires ; dans l'exemple, l'apparition du nouvel objectif, qui définit la spécificité même de l'organisation, à la fois définit son existence et la met en jeu.

– Les membres de l'organisation ne parlent pas du paradoxe, ils « tournent » dedans. C'est comme s'ils reprenaient à leur compte l'incompatibilité des deux pôles de la contradiction. Pour peu que nous ayons pu l'observer, le paradoxe n'est pas marginal, il occupe pratiquement tout le champ du discours et resurgit à de multiples moments. Il n'est pas repris comme paradoxe « personnel » mais la personne se trouve « tirée » de son côté. les paradoxes organisationnels que nous avons pu observer ici mobilisent une « énergie sémantique et pratique » hors du commun.

Références

Berger P. et Luckman T. (1986). *La construction sociale de la réalité*. Klincksieck.

Czarniawska B. (1998). *A Narrative Approach to Organization Studies*. Sage.

Eden C., Ackerman F. et Cropper S. (1992). The analysis of Cause Maps. *Journal of Management Studies*, vol. (29 : 3) : 309-324.

Grant D., Keenoy T. et Oswick Cl. (Eds). (1998). *Discourse + Organization*. Sage.

Greimas A.J. (1966). *Sémantique structurale, recherché de méthode*. Larousse.

Hassard J. (1993). Postmodernism and Organizational Analysis: an Overview. In Hassard et Parker (Eds), *Postmodernism and Organizations*. Sage.

Heritage J. (1984). *Garfinkel and Ethomethodology*. Polity Press.

Hiernaux J.P. (1977). *L'institution culturelle II, Méthode de description structurale*. Presses Universitaires de Louvain.

Jacques J-M., Gatot L. et Wallemacq A. (2002). Risk Perception and Crises Management: Landcape of Mind, Landscape of Action. In Fabbri K. and Yeroyanni M. (Eds), *Proceedings of the first EU-MEDIN Workshop on Natural and Technological Hazards* : 24-33.

Latiers M., Jacques J.-M., Bruyninckx V., Van Assche V., Laurent N. et Wallemacq A. (2003). Software, Springboard, Creativity. 10th workshop on Managerial and Organizational Cognition – Paris (Cergy-Pontoise).

Laukkanen M. (1994). Comparative Cause Mapping of Organizational Cognitions. *Organization Science*, vol. (5/3) : 322-343.

Linstead S. (2001). *Rhetoric and Organizational Control: A Framework for Analysis*. In Westwood R. et Linstead S. (Eds), *The language of Organization*. Sage.

Merleau-Ponty M. (1960). *Signes*. Gallimard (NRF).

Townley B. (1994). *Reframing Human Ressource Management: Power, Ethics and the subject at work*. Sage.

Varela F., Thompson E. et Rosch E. (1993). *L'inscription corporelle de l'esprit*. Seuil.

Varela F. (1996). *Quel savoir pour l'éthique?* La Découverte.

Wallemacq A. et Jacques J.-M. (2003). *Language and Landscapes, towards new architectures of thought*. In Linstead S. et Linstead A. (Eds). *Thinking Organization*. Routledge (A paraître).

Watzlawick P., Helmick-Beavin J. et Jackson D. (1972). *Une logique de la communication*. Seuil.

Weick K.E. (1979). *The Social Psychology of Organizing Reading*. MA, Addison – Wesley.

Westwood R. et Linstead S. (2001). *Language/Organization: An Introduction*. In Westwood R. et Linstead S. (Eds), *The language of Organization*. Sage.

Whorf B.L. (1956). *Language, Thought and Reality*. Wiley and sons.

Cadre pour la catégorisation de textes multilingues

Radwan Jalam, Jérémy Clech, Ricco Rakotomalala

Laboratoire ERIC – Université Lumière Lyon 2
5, av. Pierre Mendès-France – 69676 Bron, France
{jalam, jclech, rakotoma}@eric.univ-lyon2.fr

Abstract

In this paper, we propose an original framework for multilingual text categorization. The objective is to classify a set of texts, written in some language, using a predictive model learned from a set of texts written in a given language, called learning language. Contrary to the unilingual classical phase of text categorization, the classification phase contains two new steps : firstly identify the language of the text, and then automatically translate it into the learning language. As shown in this paper, first applications of multilingual text categorization on real data, that is over English, French and German newspapers, indicate that the approach is viable.

Résumé

Dans cet article, nous proposons un cadre pour la catégorisation de textes multilingues. L'objectif est de pouvoir, à partir d'un modèle de prédiction construit par apprentissage sur un corpus de textes rédigés dans un langue donnée, inférer sur une série de textes qui sont rédigés dans une langue quelconque. Cette phase d'inférence, par rapport à la généralisation classique, comprend deux étapes supplémentaires : la détection de la langue du texte, puis sa traduction automatique vers la langue de référence. Nos premiers résultats sur une application réelle : la catégorisation d'articles de journaux allemands, anglais et français, montrent la viabilité de l'approche.

Keywords: multilingual text categorization, language identification, machine learning, n-grams representation, CLEF collection, translation effects on text categorization.

1. Introduction

La catégorisation de texte consiste à chercher une liaison fonctionnelle entre *un ensemble de textes* et *un ensemble de catégories* (étiquettes, classes). Cette liaison fonctionnelle, que l'on appelle également *modèle de prédiction*, est estimée par un apprentissage automatique (traduction de *machine learning method*). Pour ce faire, il est nécessaire de disposer d'un ensemble de textes préalablement étiquetés, dit *ensemble d'apprentissage*, à partir duquel nous estimons les paramètres du modèle de prédiction le plus performant possible, c'est-à-dire le modèle qui produit le moins d'*erreurs* en prédiction (voir la figure 1).

Formellement, la catégorisation de texte consiste à associer une valeur booléenne à chaque paire $(d_j, c_i) \in \mathcal{D} \times \mathcal{C}$, où $\mathcal{D}$ est l'ensemble des textes et $\mathcal{C}$ est l'ensemble des catégories. La valeur V (Vrai) est alors associée au couple (d_j, c_i) si le texte d_j appartient à la classe c_i tandis que la valeur F (Faux) lui sera associée dans le cas contraire. Le but de la catégorisation de texte est de construire une procédure (modèle, classifieur) $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$ qui associe une ou plusieurs catégories à un document d_j telle que la décision donnée par cette procédure « coïncide le plus possible » avec la fonction $\breve{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{V, F\}$, la vraie fonction qui retourne pour chaque vecteur d_j une valeur c_i (Sebastiani, 2002).

Dans ce travail, nous proposons des solutions pour étendre la catégorisation de textes aux *corpus*

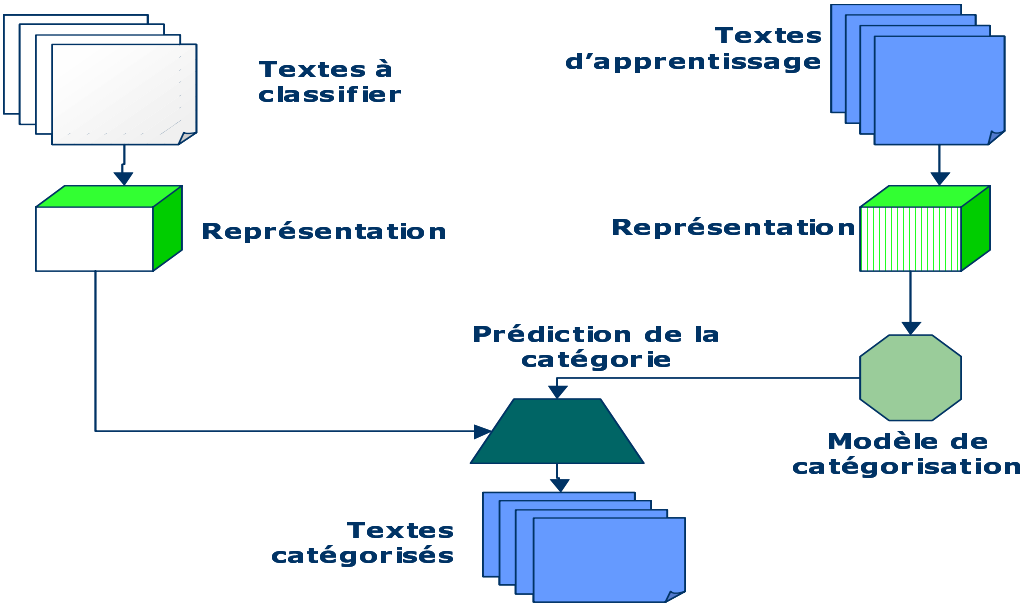


Figure 1. Schéma général pour la catégorisation de textes « monolingue ».

multilingues. Ceci introduit des contraintes supplémentaires dans la catégorisation de textes : il faut reconnaître automatiquement la langue d'un texte puis procéder à une traduction automatique. Notre approche semble naïve, mais elle a le mérite d'être :

1. la première solution automatique proposée, à notre connaissance,
2. opérationnelle, comme le montrent les premières expérimentations sur un corpus d'articles de journaux allemands, anglais et français.

La phase d'apprentissage s'effectuera, comme dans la catégorisation de texte « monolingue », à partir d'un corpus d'apprentissage étiqueté rédigé dans une langue donnée $\mathfrak{L}_{app}$. Mais l'inférence sera possible pour un texte rédigé dans une langue quelconque, dès qu'un traducteur automatique sera disponible de cette langue vers $\mathfrak{L}_{app}$. Notons que notre approche exclut les méthodes qui utilisent de manière explicite des informations spécifiques à chaque langue.

Bien entendu, à l'instar de la catégorisation de textes monolingue, le texte à classer doit appartenir au même domaine que les textes utilisés lors de l'apprentissage. On ne saurait, par exemple, essayer de classer un article scientifique à partir d'un modèle construit sur un ensemble d'apprentissage constitués d'articles de journaux à scandale.

Cet article est organisé de la manière suivante : dans la section 2, nous exposerons l'approche que nous avons pour étendre la catégorisation de textes au cas multilingue. Dans la section 3, nous mettrons en oeuvre le cadre proposé sur un exemple réel de catégorisation de journaux. Dans la section 4, nous discutons des résultats obtenus, et nous essayerons de mettre en perspective notre démarche afin de la faire évoluer. Nous conclurons alors dans la section 5.

2. Nouveau cadre pour la catégorisation multilingue

Dans le cadre de la catégorisation de textes multilingues, le processus comporte deux nouvelles exigences : le corpus de textes étiquetés utilisé pour l'apprentissage est disponible dans une langue $\mathfrak{L}_{app}$ donnée ; chaque nouveau texte à classer est dans une langue que l'on doit d'abord déterminer avant de pouvoir lui associer son étiquette.

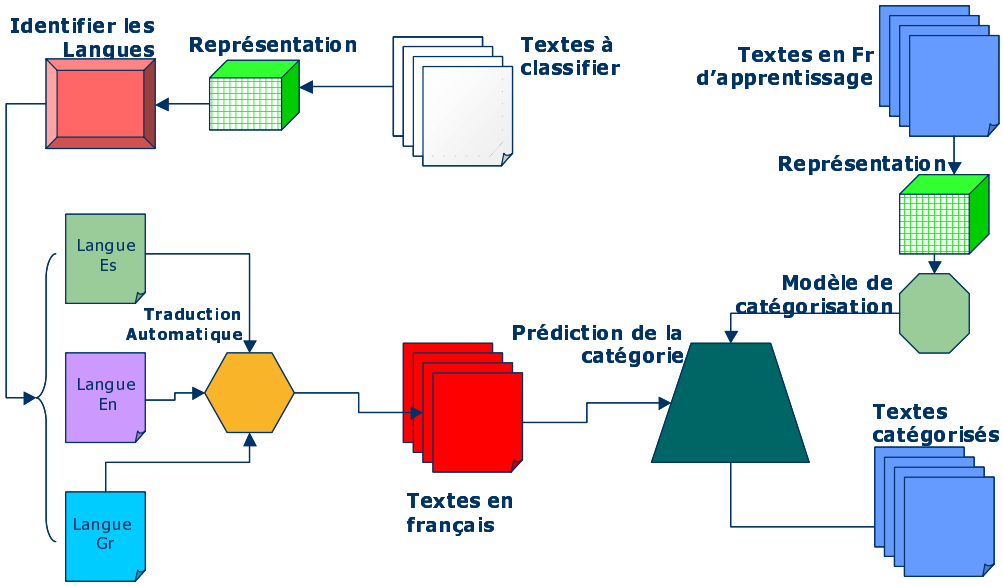


Figure 2. Catégorisation de textes multilingues : choix d'une langue d'apprentissage et utilisation d'un seul modèle, dans cette langue

Pour répondre à ces nouvelles contraintes, nous aménageons le processus de catégorisation exposé dans la figure 1. La phase d'apprentissage n'est pas modifiée, en revanche la phase de classement comporte deux étapes supplémentaires (figure 2) :

1. nous devons tout d'abord détecter la langue dans laquelle le texte d_j à classer est rédigé ;

2. si la langue est reconnue par le traducteur, il est traduit vers la langue $\mathfrak{L}_{app}$ et le texte à classer devient $d_{j_{app}}$.

Nous recherchons alors les occurrences des termes $(w_1, w_2, w_k, ..., w_{|\mathcal{T}|})$ dans $d_{j_{app}}$ afin de pouvoir appliquer le modèle prédictif et ainsi associer une catégorie c_i au texte à classer. Rappelons qu'en catégorisation de texte, chaque document d_j est transformé en un vecteur $\mathbf{d}_j = (w_{1j}, w_{2j}, w_{kj}, ..., w_{|\mathcal{T}|j})$, où $\mathcal{T}$ est l'ensemble de termes (descripteurs) qui apparaissent au moins une fois dans l'ensemble d'apprentissage. Le poids w_{kj} correspond à la contribution du termes t_k à la sémantique du texte d_j.

2.1. *Détection de la langue du texte à classer*

Il est important de détecter avec précision la langue dans laquelle le texte à classer est rédigé, car une erreur à ce niveau voue à l'échec les étapes suivantes.

Il existe deux familles d'approches dans l'identification de la langue : linguistique ou statistique. Afin de rester cohérent avec nos choix pour la catégorisation de textes, nous avons privilégié l'approche statistique. Cette approche capture automatiquement certaines régularités statistiques des langues. Nous avons utilisé les 3-grammes qui sont des séquences de 3 caractères consécutifs extraits du texte à classer (Jalam et Chauchat, 2002). Nos précédents travaux ont montré qu'un texte de longueur de 100 octets permettait d'obtenir une reconnaissance d'excellente qualité, à près de 99% (Jalam et Teytaud, 2001). Et pour des textes plus longs, la reconnaissance est parfaite.

2.2. *Traduction du texte à classer*

La traduction du texte à classer dans la langue du corpus d'apprentissage $\mathfrak{L}_{app}$ est également une étape primordiale. L'objectif ici n'est pas de produire un texte traduit retraçant fidèlement les propriétés sémantiques de l'original, mais de fournir un texte assurant une qualité de classement suffisante. Il est évident que le résultat obtenu dépendra du traducteur utilisé, une des perspectives immédiates de notre travail actuel consistera à analyser de manière approfondie le traducteur pour évaluer son efficacité lors de la catégorisation de textes.

Afin de défricher le terrain, nous avons utilisé un traducteur en ligne disponible sur Internet (http ://babelfish.altavista.com/, qui utilise la technologie Systran). Ce traducteur n'est sans doute pas le meilleur, mais comme il est publiquement disponible, nous aurons la possibilité par la suite de comparer nos résultats avec d'autres études. L'utilisation d'un meilleur traducteur ne peut qu'améliorer la catégorisation des textes.

3. Application sur les corpus CLEF

3.1. *Constitution du corpus*

La constitution de ce corpus a été un travail long et difficile ; malgré nos recherches, nous n'avons pas trouvé de corpus multilingues de textes étiquetés en classes comparables. Nous avons dû adapter pour cela les documents proposés par les organisateurs du concours CLEF (voir la page web : http://clef.iei.pi.cnr.it).

Les corpus de documents utilisés dans la campagne d'évaluation CLEF proviennent de différents journaux tels que le *Los Angeles Times* (États-Unis), *Le Monde* (France), *La Stampa* (Italie), *Der Spiegel* et *Frankfurter Rundschau* (Allemagne), d'agences de presse comme *EFE* (Espagne) ou des dépêches de l'*ATS*, l'agence télégraphique suisse, disponibles en allemand, français et italien. Les documents de ces corpus sont tous extraits de l'année 1994 et les thèmes abordés sont approximativement comparables.

Ces corpus sont destinés aux tâches de la recherche documentaire monolingues, bilingues et multilingues. Les fichiers de ces corpus sont au format SGML. Par exemple, le fichier `lemonde_19940604.sgml` concerne les articles apparus dans le journal Le Monde (LM) lors de la journée du 4 juin 1994. Chaque article de ce fichier contient des balises sgml qui décrivent son contenu (table 1). Il est facile d'y distinguer son numéro d'identification, son titre, et son corps.

La taille des corpus varie fortement entre les langues, avec des volumes plus restreints pour le français et l'italien. Le nombre de mots par article reste assez similaire (environ 130), avec une moyenne un peu plus élevée pour la collection anglaise (167). Par contre, la variabilité de cette longueur demeure assez forte (écart-type d'environ 120), sauf pour les langues espagnole, où l'écart-type est de 60, et italienne, où l'écart-type est de 97 (Savoy, 2002).

Comme on le sait, la catégorisation de textes nécessite d'avoir un échantillon d'apprentissage composé de textes associés à leurs classes. Ces exemples serviront, dans le processus d'apprentissage, à apprendre un classifieur qui sera ensuite appliqué sur les textes à classer. Une première approche peut consister à utiliser les termes-index, extraits de la balise « *subjects* », associés à chaque article comme classes d'appartenance. En général, l'indexation est une opération qui *décrit* et *caractérise* un document, ou un fragment de document, en repérant les thèmes présents dans ce document (Afnor, 1993). Malheureusement, les résultats de nos expérimentations utilisant les termes-index, sur les corpus français et allemand, sont médiocres : le taux d'erreur est proche de celui du hasard.

```
<DOC>
<DOCNO>LEMONDE94-000386-19940604</DOCNO>
<DOCID>LEMONDE94-000386-19940604</DOCID>
<ACCOUNT>339369</ACCOUNT>
<GENRE>RECTIF</GENRE>
<DATE>19940604</DATE>
<LMDOC>MHB</LMDOC>
<FAB>06031011</FAB>
<SUBJECTS>DEMENTI,DESSIN</SUBJECTS>
<GO21>PUBLICATION</GO21>
<NAMES>SERGUEI</NAMES>
<PUM1>QUO</PUM1>
<REFERENCE1>2-002-06</REFERENCE1>
<SEC1>IDE</SEC1>
<PAGE>13</PAGE>
<TITLE>Serguei précise</TITLE>
<TIO1>PAS DE PANIQUE A BORD</TIO1>
<TEXT>Un lecteur s'est étonné de constater qu'une publication satirique d'extrême droite, Pas de panique à bord, citait, parmi les noms de ses collaborateurs, celui du dessinateur Sergueï. Etonnement encore plus grand _ pour ne pas dire plus _ de Sergueï, celui que nos lecteurs connaissent bien et qui ne saurait être le même que son homonyme de Pas de panique à bord, s'il existe. &gt ;</TEXT>
</DOC>
```


Tableau 1. Exemple extrait de la collection CLEF et qui montre un article du journal Le Monde publié le 4 juin 1994.

Il est difficile d'apprendre, si l'on utilise les termes-index proposés en tant que classes, pour différentes raisons :

– Les termes-index associés aux articles français et allemand sont trop nombreux, plus de 16 200 termes (et donc classes) pour le corpus français et plus de 32000 termes pour le corpus SDA allemand. Il n'y a pas eu de règle d'indexation basée sur des vocabulaires contrôlés et beaucoup de termes sont des synonymes (mort, morts ; France, fr ; German, gr).
– Les termes-index proposés ne représentent vraiment pas les thèmes abordés dans les articles ; par exemple, on associe à la dépêche de la table 1 le mot clé « *dessin* » alors qu'elle parle d'un démenti.
– Les termes-index dans les dépêches de l'agence télégraphique suisse (allemand et français) ne sont pas séparés par des signes de ponctuation ainsi, il est difficile d'extraire les termes composés comme *« conseil de sécurité »*. Le corpus de Los Angeles Times, à la différence des autres corpus, ne fournit pas de tout de termes-index, ces termes-index aident normalement à décrire le contenu d'un article.

Dans nos expérimentations, nous choisissons donc de considérer les thèmes proposés dans la campagne CLEF 2002 comme classes à prédire. Ces dernières sont très variées ; on trouve par exemple : *« U.N. sanctions against Iraq »*, *« Conflict in Palestine »* ou *« Leaning Tower of Pisa »*. Nous avons travaillé sur trois corpus de langues anglaise, française et allemande : le *Los Angeles Times* (LAT), *Le Monde* (LM) et *l'agence télégraphique suisse* (SDA). Les thèmes utilisés dans nos évaluations sont décrits dans la table 2.

3.2. *Représentation des textes*

La représentation des textes est une étape critique. Nous avons choisi d'utiliser les mots et les 3-, 4- et 5-grammes. Nous avons appliqué notre algorithme de sélection de termes χ^2_{multi},

Classes	CLEF id	CLEF topic	#LAT	#LM	#SDA
C_1	92	U.N. sanctions against Iraq	27	24	23
C_2	95	Conflict in Palestine	96	89	66
C_3	103	Conflict of Interests in Italy	10	24	70
C_4	108	Southern Yemen Secession	18	19	63
C_5	119	Destruction of Ukrainian nuclear weapons	54	33	55
C_6	122	North American car industry	27	23	6
C_7	124	Common foreign and security policy (CFSP)	32	48	24
C_8	131	Intellectual Property Rights	40	43	43
C_9	133	German Armed Forces Out-of-area	10	21	17
C_{10}	140	Mobile phones	70	95	23
Total			**384**	**419**	**390**

Tableau 2. Description des catégories utilisées.

présenté dans (Clech *et al.*, 2003), en sélectionnant 100 mots avec pour seuls prétraitements l'uniformisation de la casse et la suppression des mots-outils (*Stop Words*). En outre, nous avons sélectionnés 200 3-, 4- et 5-grammes avec pour seul prétraitement l'uniformisation de la casse. Nous avons choisi de sélectionner moins de mots que de n-grammes puisqu'un mot est composé de plusieurs n-grammes ; après plusieurs expériences, le choix de 100 mots et 200 n-grammes apparaît comme un bon compromis conservant la structure informationnelle du corpus.

L'étude des termes sélectionnés révèle la présence importante de noms propres, tels les noms des pays ou de leurs ressortissants, ou encore les noms de personnalités. L'étude indique également certaines difficultés de traduction. Par exemple, l'expression française « téléphone portable » est traduite en anglais par « *portable phone* » ce qui n'a pas le sens voulu. Les noms propres ne sont pas non plus épargnés par des difficultés de traduction ; ainsi le terme français « Koweït » est laissé tel quel au lieu d'être traduit par le terme « Kuwait ».

3.3. *Algorithmes d'apprentissage*

Dans notre application, nous utilisons deux méthodes renvoyant à des paradigmes d'apprentissage très différents :

– une méthode arborescente : C4.5 (Quinlan, 1993). Cette algorithme glouton a la particularité de sélectionner les variables et effectue des découpages parallèles aux axes ;
– une méthode d'estimation des probabilités locales : les 3 plus proches voisins (3-ppv) (Aha *et al.*, 1991). Cet algorithme est sensible aux variables non pertinentes ainsi qu'aux espaces de représentations creux (Mitchell, 1997).

3.4. *Reconnaissance de la langue*

Dans nos expérimentation, nous avons utilisé la distance de χ^2 pour identifier la langue de texte parmi les trois langues français, anglais et allemand. Nos nouveaux tests confirment nos résultats précédents (Jalamet et Teytaud, 2001) : pour des textes de taille égale ou supérieure à 100 caractères le taux de reconnaissance de la langue est de 100%.

3.5. *Catégorisation des articles*

Afin de pouvoir évaluer notre processus de catégorisation de textes dans un corpus multilingue, nous avons effectué plusieurs expériences de catégorisation. Pour chacune d'elles, nous avons

	10-V.C. LAT		10-V.C. LM		10-V.C. SDA	
taux d'erreur	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**	**3-NN**
100 mots	16%	8%	24%	5%	15%	3%
200 3-grammes	23%	9%	20%	9%	11%	2%
200 4-grammes	16%	7%	16%	5%	8%	2%
200 5-grammes	14%	7%	16%	5%	9%	2%

Tableau 3. Taux d'erreur, en validation croisée (V.C.), pour les articles de journaux dans leur langue originelle.

	LM (An)		SDA (An)	
Taux d'erreur	**C4-5**	**3-NN**	**C4-5**	**3-NN**
100 mots	25%	6%	12%	3%
200 3-grammes	16%	6%	9%	3%
200 4-grammes	12%	6%	9%	2%
200 5-grammes	13%	5%	12%	3%

Tableau 4. Taux d'erreur, en validation croisée, pour les articles de journaux traduits en anglais. LM (An) désigne le corpus français Le Monde (LM) traduit en anglais (An). SDA (An) désigne le corpus allemand de l'agence télégraphique suisse (SDA) traduit en anglais (An).

évalué le taux d'erreur pour toutes les configurations de représentations des textes (100 mots, 200 3-grammes, 4-grammes et 5-grammes) et des modèles utilisés (C4.5 et les 3 plus proches voisins).

Usuellement, en catégorisation de textes, un document peut appartenir à n classes. Cependant, dans l'application présentée ici (voir le tableau 2), chaque document n'appartient qu'à une classe, et une seule. Il en résulte que les rappel et précision micro-moyennes sont identiques et égaux au taux de succès ; pour cette raison nous allons aussi utiliser le taux d'erreur pour mesurer les performances.

Nous présentons tout d'abord les résultats en catégorisation monolingue (tableau 3), ils ont servi de référence pour juger de la qualité de ceux obtenus dans un contexte multilingue.

3.5.1. *Catégorisation monolingue*

Afin d'évaluer le niveau intrinsèque de difficulté de catégorisation des articles, nous avons mesuré l'erreur de classement pour nos 3 corpus (LAT, LM et SDA) dans leur langue d'origine.

Les résultats (par *10-validation croisée*) sont présentés dans le tableau 3 ; nous en dégageons 4 principaux résultats :

– Il y a un apprentissage effectif puisque le taux d'erreur est largement inférieur aux taux d'erreur du classifieur par défaut ;
– Il y a un avantage important pour la méthode du 3-ppv (un écart supérieur à 10 points) par rapport à C4.5 qui souffre de la fragmentation des données ;
– Le bon apprentissage du 3-ppv laisse penser que les termes sélectionnés sont pertinents ;
– Le corpus allemand est plus facile à catégoriser que les deux autres.

(a) représentation avec 100 mots

	Appris et appliqué sur LM (Fr)		**Appris sur LAT (An) et appliqué sur**			
	LM (Fr)		**LM (Fr)**		**LM (An)**	
	3-NN	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	95%	76%	78%	12%	89%	60%
ρ^{M}	94%	68%	78%	14%	92%	55%
π^{M}	92%	71%	80%	10%	88%	52%

(b) représentation avec 200 4-grammes

	Appris et appliqué sur LM (Fr)		**Appris sur LAT (An) et appliqué sur**			
	LM (Fr)		**LM (Fr)**		**LM (An)**	
	3-NN	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	95%	84%	86%	55%	90%	68%
ρ^{M}	96%	80%	81%	43%	89%	59%
π^{M}	95%	79%	90%	32%	91%	50%

Tableau 5. Précision et Rappel (micro et macro-moyen), pour le corpus Le Monde (LM) représenté par 100 mots (table a) et pour 200 4-grammes (table b). Les deux tables montrent les performances obtenues en validation croisée. On applique le modèle LM (Fr) sur des textes écrits en français et les résultats obtenus sont comparés avec ceux du modèle LAT anglais sur le corpus « LM écrit en français » et sur le corpus « LM traduit en anglais ».

3.5.2. *Catégorisation multilingue*

Comme décrite précédemment, la catégorisation multilingue nécessite des étapes intermédiaires complémentaires, chacune pouvant générer du biais pour le processus même de la catégorisation. En effet, si l'étape de traduction est de qualité médiocre, l'apprentissage sera difficile et les résultats de la catégorisation seront de mauvaise qualité. Par ailleurs, comme nous avons dit dans la section 3.2, les noms propres sont très présents dans les mots sélectionnés, et de ce fait nous pourrions nous demander si de bons résultats de catégorisation ne seraient pas simplement dus à ces noms propres. Ainsi, dans l'objectif de valider notre processus de catégorisation multilingue, nous évaluons d'abord l'effet de la traduction, puis l'effet potentiel des noms propres et enfin la capacité même du modèle à être appliqué sur des textes traduits.

Les effets du traducteur Pour mesurer l'effet du traducteur sur le contenu informationnel des documents, nous avons traduit vers l'anglais le corpus français LM et le corpus allemand SDA. Nous avons appliqué le schéma d'apprentissage monolingue (voir la figure 1) et évalué l'erreur en validation croisée (tableau 4). Les résultats montrent que le contenu informationnel des corpus, du moins ce qui est nécessaire à la catégorisation, est très peu dégradé par la traduction. En effet, les différences des taux d'erreurs obtenus après traduction (tableau 4) comparées à ceux obtenus dans la langue d'origine (tableau 3) ne sont pas significatives.

Les effets de noms propres Pour évaluer comment les noms propres peuvent influencer les résultats de l'étape d'apprentissage, nous avons estimé un modèle à partir du corpus LAT dans sa langue d'origine (anglais) que nous appliquons directement sur les corpus LM et SDA laissés

dans leur langue d'origine (respectivement français et allemand). Nous supposions que, si les résultats ne sont pas significativement différents de ceux obtenus après traduction, alors, la contribution des noms propres lors de l'étape de catégorisation est supérieure à la contribution des mots communs traduits.

Enfin, nous avons étudié les résultats de catégorisation après traduction en anglais des corpus à classer (LM et SDA) en appliquant le modèle élaboré à partir du corpus du LAT en anglais.

Nous présentons seulement les résultats concernant la représentation utilisant 100 mots et celle utilisant 200 4-grammes. Le tableau 5 regroupe les résultats obtenus pour le corpus LM et le tableau 6 ceux de SDA. Nous en dégageons 3 résultats principaux :

– Le premier concerne la viabilité de notre approche : même si le taux d'erreur s'accroît quand on passe d'un apprentissage sur les traductions anglaise au lieu des textes originaux, la qualité de prédiction surpasse largement celle du classifieur par défaut.
– Le second concerne la faible variabilité des résultats obtenus en fonction des représentations de texte utilisées (mots ou n-grammes) : on ne peut pas dire si l'une est plus robuste que l'autre.
– La troisième concerne le faible biais introduit par les noms propres ; les tableaux 5 et 6, montrent que l'écart entre les résultats avant et après traduction sont suffisamment significatifs : nous observons pour le modèle 3-ppv un saut de plus de 10 points pour le corpus LM (Tableau 5) et un saut de plus de 70 points pour le corpus SDA (Tableau 6) ; rappelons que nous supposions que, si les résultats du modèle estimé pour l'anglais mais appliqué à des textes en français ou allemand, n'étaient pas significativement inférieurs de ceux obtenus sur ces textes traduits, alors, la contribution des noms propres lors de l'étape de catégorisation serait supérieure à la contribution des mots communs traduits.

Par ailleurs, le taux d'erreur de C4.5 s'explique largement par la difficulté de prédire les classes c_3, c_6 et c_9 dont les taux de rappels et de précisions sont nuls. Pour c_3 et c_9 cela est dû au seuil d'élagage (fixé à 10) correspondant au nombre de textes du corpus d'apprentissage (LAT) composant ces classes (voir le tableau 2). Pour c_6, C4.5 produit une seule règle, basée sur la présence du mot 'auto', qui provient des expressions *« auto manufacter*s » ou *« auto shows »*, dans le corpus du LAT ; or les expressions françaises *« salon de l'auto »* et *« industrie automobile »* des articles du Monde (LM) sont respectivement traduites par *« car show »* et *« car industries »* : le terme *« auto »* étant traduit par *« car »*, le terme *« auto »* est absent des traductions de LM ; C4.5 ne peut donc appliquer son (unique) règle apprise.

4. Discussion

Les résultats obtenus par nos modèles sur notre corpus sont des résultats encourageants. Cette section propose de discuter les différents choix effectués pour chacune des étapes du processus afin de moduler la signification de nos résultats.

La première étape du processus consiste à définir une représentation du corpus par des termes. Nous avons choisi ici la représentation basée sur les mots et celle basée sur les n-grammes. Ce dernier choix était motivé par la capacité des n-grammes à capturer aisément les structures informationnelles basiques en s'affranchissant des problèmes de séparation des mots, de coquilles et tout autre aspect linguistique. Nos expériences n'ont pas montré une différentiation marquée entre les résultats issus d'une sélection de mots et d'une sélection de n-grammes. Nous attribuons ces similarités à la qualité contrôlée des corpus. En effet, ces derniers sont destinés à la presse écrite, qui est exigeante envers les fautes d'orthographes et coquilles.

(a) représentation avec 100 mots

	classifieur SDA (Ger) appliqué sur		**classifieur LAT (An) appliqué sur**			
	SDA (Ger)		**SDA (Ger)**		**SDA (An)**	
	3-NN	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	97%	85%	20%	17%	97%	56%
ρ^{M}	93%	77%	13%	14%	95%	50%
π^{M}	95%	70%	12%	25%	94%	62%

(b) représentation avec 200 4-grammes

	classifieur SDA (Ger) appliqué sur		**classifieur LAT (An) appliqué sur**			
	SDA (Ger)		**SDA (Ger)**		**SDA (An)**	
	3-NN	**C4.5**	**3-NN**	**C4.5**	**3-NN**	**C4.5**
$\rho^{\mu} = \pi^{\mu}$	98%	92%	21%	17%	97%	54%
ρ^{M}	94%	80%	14%	14%	97%	52%
π^{M}	94%	78%	12%	25%	96%	54%

Tableau 6. Précision et Rappel (micro et macro-moyen) pour le corpus allemand SDA représenté par 100 mots (tableau a) et pour 200 4-grammes (tableau b). Les deux tableaux montrent les performances obtenues en validation croisée : On applique le modèle SDA (Ger) sur des textes écrits en allemand et les résultats obtenus sont comparés avec ceux du modèle LAT anglais sur le corpus « SDA écrit en allemand » et sur le corpus « SDA traduit en anglais ».

La deuxième étape consiste en la sélection des termes. Les coprésences des mots sélectionnés à partir du corpus du LAT et de celui du LM traduit en anglais se situent au niveau de 50 %. La moitié d'entre-eux est constituée de mots communs. Nous obtenons des résultats similaires lors de la comparaison des termes sélectionnés à partir du corpus du LAT et de celui du SDA traduit en anglais. Ceci montre la similitude informationnelle de nos 3 corpus. Par ailleurs, la forte quantité de noms propres n'est pas un problème en lui-même, et nous avons vu que son apport était faible. Cependant, lorsque le traducteur ne connaît pas un terme il le laisse tel quel. De plus, les noms propres ont une faible variabilité d'écriture dans les différentes langues. Ainsi, nous pensons que les noms propres empêchent l'évaluation complète de l'effet de la traduction.

Enfin, la sélection des termes est qualitativement intéressante puisqu'elle permet une séparabilité aisée de nos 10 classes. Là encore, nous nous interrogeons sur la constitution de notre corpus. Le corpus CLEF contient 96 000 articles. De ces 96 000, seuls 8 000 ont été assigné à 50 sujets (classes). Pour améliorer notre corpus et confirmer nos résultats, nous envisageons la généralisation en travaillant sur l'ensemble des 96 000 textes (8 000 assignés à des classes définies et les 88 000 restants assignés à une classe « autre ») ceci rendra plus difficile la séparabilité des sujets.

La dernière étape concerne l'apprentissage ; elle a montré les bons résultats du 3-ppv sur ces corpus. Ce résultat est en opposition avec la difficulté de prédiction des k-ppv dans un espace creux et/ou avec des variables non pertinentes. Nous attribuons donc ces bons résultats des 3-ppv à la qualité de notre espace de représentation (bon choix des descripteurs par notre méthode du χ^2_{multi} multivarié). Enfin, le C4.5 donne des résultats convenables mais est moins performant

que le 3-ppv. Comme nous l'avons vu, cela est du au faible effectifs de certaines classes et au paramétrage de la méthode.

5. Conclusion

L'objet de cet article est la définition d'un processus pour la catégorisation multilingue. Nous avons introduit deux nouvelles étapes par rapport au processus monolingue : la détection de la langue du texte à catégoriser et sa traduction dans la langue du corpus d'apprentissage. Nous avons illustré notre procédé par une application sur des corpus réels de journaux écrits en 3 langues (anglaise, française et allemande). Nous avons décrit chaque étape de notre processus et présenté les résultats de nos expériences. Nous concluons à la l'efficacité de notre approche.

Nous envisageons de perfectionner notre cadre pour la catégorisation de textes multilingues en proposant un schéma plus général dans lequel nous fusionnerons des corpus d'apprentissage traduits en une langue commune d'apprentissage. Nous espérons ainsi que les particularités propres à chaque langue ne seront plus retenues par les modèles estimés.

Références

Afnor (1993). Information et documentation. Principes généraux pour l'indexation des documents. NF Z 47-102.

Aha D.W., Kibler D. et Albert M.K. (1991). Instance-based learning algorithms. *Machine Learning*, vol. (6) : 37-66.

Clech J., Rakotomalala R. et Jalam R. (2003). Séléction multivariée de termes. In *Société Française de Statistique* : 933-936.

Jalam R. et Chauchat J.-H. (2002). Pourquoi les n-grammes permettent de classer des textes ? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques. In *Actes des JADT 2002*, vol. (1) : 381-390.

Jalam R. et Teytaud O. (2001). Kernel based text categorization. In *Proceedings of IJCNN-01, 12th International Joint Conference on Neural Networks* : 1891-1896.

Mitchell T.M. (1997). *Machine Learning*. Computer Science. McGraw-Hill.

Quinlan J. (1993). *C4.5 : Programs for Machine Learning*. San Mateo, CA.

Savoy J. (2002). Report on clef-2002 experiements : Combining multiple sources of evidence. Results of the CLEF-2002, cross-language evaluation forum. University of Neuchatel.

Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, vol. (34/1) : 1-47.

Recherche de structures latentes dans des partitions de « textes » de 2 à K classes

Michèle Jardino

LIMSI – CNRS – BP 133 – 91403 Orsay – France
jardino@limsi.fr

Abstract

We have developed an original learning method in order to extract latent structures in raw texts. The induced structure is a data-driven tree which can be unbalanced. It has been obtained from successive partitions of the texts in classes, with an incremental number of classes ranging from 2 to k; each quasi-optimal partition has been performed with an adaptation of the k-means clustering. The paths of the texts in the successive partitions are the edges of an oriented graph whose nodes are the clusters. The study of the paths shows that some of the clusters remain identical in the successive partitions so that a tree can be extracted from the graph, by merging nodes and clipping edges. A corpus of 1100 touring information leaflets has been used to illustrate this method.

Résumé

Nous proposons une méthode automatique pour extraire une structure, si elle existe, dans des textes ou des ensembles de textes. Cette structure prend la forme d'un arbre de profondeurs et de ramifications variables qui sont déterminées par les données. Pour cela nous réalisons des partitions successives quasi-optimales des textes à l'aide d'un algorithme de classification non supervisée de type « centres de gravité mobiles ». Nous étudions ensuite les chemins effectués par les textes dans ces partitions successives. Ces chemins forment un treillis dans lequel on peut extraire par fusion et élagage un arbre si certaines classes restent stables dans les partitions successives. Nous illustrons cette recherche sur un corpus constitué d'un ensemble de 1100 fiches touristiques découpées en fragments de textes.

Mots-clés : structure latente de textes, arbre non équilibré, classification non supervisée, algorithme centres de gravité mobiles, théorie de l'information, entropie, recherche aléatoire.

1. Introduction

1.1. *Problème*

Étant donné un ensemble de « textes », par exemple des fragments de textes ou des pages Web, comment en faire une partition thématique non supervisée cohérente sans connaître a priori le nombre de classes et en faisant émerger une structure hiérarchique si elle existe ? Nous notons « textes » entre guillemets, car notre méthode est générique et peut être appliquée à tout type de données qui peuvent être représentées par des vecteurs (voir section 2.1).

1.2. *État de l'art*

L'abondance des textes électroniques disponibles aujourd'hui a donné un regain d'intérêt aux méthodes de classification non supervisées : ces méthodes automatiques sont adaptées à la gestion de grands flots de données. Deux types de partitionnement non supervisés sont généralement utilisés : la classification hiérarchique (ascendante ou descendante) qui permet de générer un arbre complet allant du regroupement de tous les éléments à classer dans une seule classe, à la racine de l'arbre, à la répartition de chaque élément dans sa propre classe, aux feuilles de l'arbre. L'autre méthode est le partitionnement direct en un nombre de classes spécifié à l'avance. Ces méthodes ont été développées dans différents logiciels commerciali-

sés, et sont disponibles gratuitement dans le logiciel libre, R (The R Project for Statistical Computing, http ://www.r-project.org/). Elles sont exposées dans de nombreux ouvrages, nous citerons Duda et Hart (1973), Kaufman et Rousseeuw (1990), Lebart *et al.* (1995). Dans ces méthodes, le choix du nombre de classes à choisir est un problème complexe, que ce soit pour la coupure de l'arbre généré par la classification hiérarchique ou pour le partitionnement direct. Il est généralement résolu soit par des connaissances a priori sur les données soit par une combinaison de méthodes de classification et d'analyse factorielle (Lebart *et al.*, 1995).

Dans tous les cas, cette coupure ne prend pas en compte la structuration réelle des données qui peuvent être éventuellement représentées par un arbre déséquilibré, avec des degrés de ramifications différents, comme par exemple pour les ontologies. Nous proposons ici une méthode pour trouver un tel arbre, si il existe, dans des données.

Proposition :

On recherche des classes qui restent stables d'une partition à une autre, c'est-à-dire des classes qui conservent les mêmes « textes ». Nous générons le treillis des relations observées entre classes de textes dans des partitions successives de 2 à K classes. L'analyse de ce treillis permet d'extraire, quand elle existe, une structure particulière sous forme d'arbre, sans connaissance a priori de cette structure, par fusion et élagage.

Nous avons étudié deux indices pour caractériser la stabilité des classes : le critère de partitionnement lui-même (Duda et Hart, 1973), et le nombre de chemins observés entre les différentes partitions.

Cette méthode peut paraître gourmande en temps de calcul car elle nécessite la recherche d'un ensemble de partitions successives. Elle bénéficie en fait d'un algorithme de classification rapide qui la rend aujourd'hui tout à fait accessible même pour traiter de grandes masses de données.

Nous décrirons d'abord comment nous représentons les textes et le corpus que nous avons utilisé pour tester cette méthode. Puis nous expliquerons notre algorithme de classification non supervisée. Enfin nous montrerons comment nous extrayons un arbre du treillis obtenu en illustrant par des graphes nos résultats. Nous mettrons ainsi en évidence les structures latentes dans ce corpus.

2. Représentations des textes

2.1. *Représentation vectorielle des textes dans l'espace des mots*

Pour représenter les textes, nous utilisons l'approche « sac de mots », largement utilisée en recherche d'information : un texte est décrit par l'ensemble des mots qu'il contient sans tenir compte de leur ordre. Ce qui est pris en compte c'est de savoir que les mots sont ensemble dans le texte, ce qui correspond à une notion de cooccurrence élargie : deux textes seront proches s'ils ont en commun un grand nombre de mots. Cette proximité peut se mesurer de manière binaire (présence ou absence) ou de manière plus précise en tenant compte de la fréquence des mots dans les textes. C'est le choix que nous avons fait : chaque texte est représenté par un vecteur dont les composantes sont les fréquences des mots normées par la longueur du texte, de sorte que la somme des composantes de chaque vecteur est égale à 1. Cette normalisation a été choisie car elle permet de représenter par un même vecteur les textes ayant une même distribution de mots comme par exemple un texte qui serait la concaténation d'un même texte répété plusieurs fois. Cette représentation est indépendante de la longueur du texte, elle correspond à la notion de profil utilisée en statistique (Lebart *et al.*, 1995).

En supposant qu'il y a T textes qui comportent au total V mots de vocabulaire différents, les textes sont indexés par i (i variant de 1 à T) et notés t_i, les mots par j (j variant de 1 à V) et notés m_j. Différentes notations du profil sont utilisées en statistique (Saporta, 1990 ; Lebart *et al.*, 1995), nous avons choisi de le noter f_i^j et d'utiliser une variante pour les fréquences marginales des textes, en remplaçant $f_{i.}$ par l_i car cette grandeur correspond à la longueur des textes et est plus facile à lire. Donc chaque texte t_i est représenté par un vecteur { f_i^j } dont les éléments f_i^j sont les fréquences relatives des mots dans le texte. On obtient ainsi une matrice de T lignes et V colonnes, chaque ligne correspond au profil d'un texte.

2.2. *Corpus d'étude*

Il s'agit de 1 100 fiches de descriptions de sites, de circuits ou de séjours touristiques. Afin de faciliter la recherche d'information dans ces fiches, par exemple pour répondre à la question « je cherche un séjour avec vue sur lagon », nous avons cherché à extraire automatiquement une structure de l'ensemble du corpus. Pour cela nous avons découpé chaque fiche en phrasettes puis regroupées celles-ci en classes thématiques selon la méthode décrite plus loin. Ce terme de phrasette au lieu de phrase a été choisi car la description des séjours dans les fiches est faite sans verbe et les segments de textes obtenus ne sont pas syntaxiquement des phrases. Les textes comprennent néanmoins des signes de ponctuation, ce qui nous a permis de les découper : chaque phrasette est un segment de texte compris entre deux points ou entre un début de description et un point. Finalement 4 700 phrasettes différentes ont été découpées dans ce corpus, à chaque phrasette est associé un vecteur. Le nombre total de mots de ce corpus est d'environ 120 000, la taille du vocabulaire est de 6 300 mots.

3. Partitions des textes

Nous présentons d'abord la méthode de partitionnement non supervisée en un nombre de classes prédéfini, ensuite nous montrons comment nous enchaînons les différentes partitions.

3.1. *Création d'une partition en k classes*

Nous disposons de T textes que nous voulons regrouper en classes homogènes bien séparées les unes des autres avec un algorithme qui permette de traiter un grand nombre de données en un temps raisonnable. Nous avons choisi la méthode de classification autour des centres mobiles (Lebart *et al.*, 1995) avec un critère de distance entropique et une recherche aléatoire de la meilleure classification (Jardino, 2000).

3.1.1. *Représentation des classes*

Chaque classe de textes est notée C_C (c varie de 1 à k) et est représentée par le vecteur associé au barycentre des textes de la classe. Ce vecteur correspond au profil de la classe de textes et sera noté f_c^j. On tient compte ainsi à la fois de la distribution des mots dans les textes (du profil des textes) et de la longueur de ces textes. L'expression du barycentre donne :

$$l_c \times f_c^j = \sum_{t_i \in C_c} l_i \times f_i^j$$

où l_c est la somme des longueurs des textes de la classe C_c et f_c^j la proportion du mot m_j dans la classe de textes C_c.

3.1.2. *Entropie, critère de classification*

Nous avons choisi comme critère de classification, l'entropie qui est une grandeur issue de la théorie de l'information (Cover et Thomas, 1991). Ce critère permet de déterminer la quantité d'information contenue dans les textes. Dans notre modèle de représentation « sac de mots », il est équivalent au nombre moyen de mots qui permettent de prédire un texte. L'entropie des

textes non classés est :

$$H(T) = -(1/f)\sum_{i=1}^{T}\sum_{j=1}^{V} f_i^j \times \log(f_i^j / l_i)$$

où f est le nombre total de mots du corpus.

Si chaque texte contient tous les mots de manière équiprobable, on a H(T) = - log(V) et on ne peut pas distinguer les textes les uns des autres. Si chaque texte est déterminé de manière unique par les mots qu'il contient on a H(T) = 0. Les valeurs d'entropie mesurées sur des corpus de textes réels sont comprises entre ces deux extrêmes.

L'entropie des textes regroupés en k classes représentées par leurs centres de gravité est :

$$H(k) = -(1/f)\sum_{c=1}^{k}\sum_{j=1}^{V} f_c^j \times \log(f_c^j / l_c)$$

On montre que l'entropie des textes regroupés H(k), quelque soit le regroupement, est plus grande que l'entropie des textes H(T) sauf dans le cas où le nombre k est égal au nombre de textes T et où chaque texte est seul dans sa classe, alors les deux entropies sont égales (Jardino, 2000).

Ce critère est intrinsèque aux centres de gravité des classes, et ne nécessite que de connaître les positions des centres de gravité et non pas leurs positions relatives, ce qui réduit les temps de calcul en comparaison de ceux obtenus avec un critère de distance ou de similarité.

3.1.3. *Algorithme de classification*

Pour une valeur de k donnée, la classification automatique cherche parmi les environ T^k configurations possibles (Jardino, 2000), celle qui minimise H(k).

Au lieu de rechercher d'emblée une classification optimale, nous avons choisi de rechercher au fur et à mesure une classification simplement meilleure que la précédente en cherchant aléatoirement les nouvelles configurations. En effet la recherche d'une solution optimale par la méthode du gradient conduit inéluctablement vers un optimum local, ce que nous voulons éviter. Cette recherche aléatoire a été inspirée par la méthode du recuit simulé (Jardino, 1993). L'objectif de cette méthode d'optimisation est d'éviter les minimums locaux en permettant d'accepter, dans un créneau fixé, des solutions qui vont en sens inverse du critère d'optimisation. En fait, comme l'ont remarqué différents utilisateurs, cette méthode bénéficie essentiellement de la recherche aléatoire des solutions et c'est finalement cet aspect que nous avons retenu. Ainsi nous obtenons des partitions qui sont quasi-indépendantes des conditions initiales.

En bref, l'algorithme de classification est le suivant :

– On choisit une partition initiale : à chaque texte est attribuée une classe puis les centres de gravité des classes sont calculés ainsi que l'entropie H(k) correspondante.

– De manière itérative, un texte est choisi au hasard et une nouvelle classe lui est attribuée également au hasard. Les centres de gravité de la classe initiale et de la nouvelle classe sont recalculés et la variation d'entropie associée s'en déduit. Si l'entropie décroît, le texte est affecté à la nouvelle classe, si elle croît, le texte reste dans la classe initiale. Le processus est réitéré jusqu'à ce qu'il n'y ait plus de variation d'entropie.

– En sortie nous disposons du classement final.

3.2. *Création de partitions de 2 à K classes*

Nous effectuons des partitions successives de 2 à K classes. La première partition est initialisée avec tous les textes regroupés dans une seule classe, les suivantes avec les résultats de classification obtenus dans la partition précédente : la partition 3, P3, est initialisée avec le classement obtenu lors de la partition P2. Les dénominations P2, P3, …, Pk, …, PK sont réservées aux partitions optimales en 2, 3,…, k, …, K classes.

Cette initialisation introduit un léger biais du fait que les partitions ne sont pas complètement optimales mais elle permet de mettre en évidence facilement les classes stables, car celles-ci conservent leur indice d'une partition à l'autre. Elle n'est pas indispensable, car comme nous le verrons ci-dessous, c'est la connaissance du chemin des textes dans les différentes partitions qui est importante.

3.3. *Partitions réalisées*

Des partitions de 2 à 10 classes ont été effectuées sur le corpus. Le tableau 1 suivant, rassemble des données et des résultats de partitionnement du corpus.

Nombre de textes, T	4 700 phrasettes
Taille du vocabulaire, V	6 300 mots
Nombre total de mots	120 000 mots
Entropie maximale	377 mots
Entropie minimale	25 mots
Entropie H(2)	295 mots
Temps de calcul	2 s
Entropie H(10)	162 mots
Temps de calcul	28 s

Tableau 1. Informations sur les partitions effectuées sur le corpus

Les entropies sont calculées en nombre moyen de mots qui prédisent les classes, cette valeur est calculée en prenant la valeur exponentielle de l'entropie, elle correspond à la notion de perplexité (Jelinek, 1988). L'entropie maximale est celle obtenue en rassemblant tous les textes dans une seule classe, on a donc un seul grand texte qui peut être prédit par 377 mots en moyenne. Cette dernière valeur est beaucoup plus petite que la taille du vocabulaire car elle tient compte de la redondance des mots. L'entropie minimale, 25 mots, correspond à la répartition des phrasettes dans 4 700 classes (une phrasette par classe).

Les temps de calcul sont courts, 2s pour une partition en 2 classes, 28s pour une partition en 10 classes.

4. Recherche d'un arbre dans le treillis

4.1. *Treillis*

Les partitions successives des textes de 2 à K classes fournissent un treillis à K niveaux. Il y a au maximum K ! branches qui connectent les classes dans les niveaux successifs. Un exemple est montré dans le tableau suivant, tableau 2, où sont recensés les chemins qui connectent le plus grand nombre de phrasettes dans les partitions de 2 à 10 classes. Chaque classe est représentée par un numéro allant de 1 à k pour chaque partition en k classes et chaque chemin est représenté par une suite de 9 chiffres, le premier chiffre de la suite varie entre 1 et 2, le deuxième entre 2 et 3 …

chemins	nombre de phrasettes ayant empruntées ces chemins
1 1 1 1 1 1 1 1 1	737
2 2 2 2 2 2 2 2 2	609
1 1 4 4 4 4 4 4 4	536
2 2 2 2 6 6 6 6 6	411
2 3 3 3 3 3 3 3 3	343
2 3 3 5 5 7 7 7 7	264
2 2 2 2 2 2 2 9 9	216
1 3 3 5 5 7 7 7 7	146
1 3 4 5 5 5 5 5 5	136
2 2 2 2 6 6 6 6 9	110

Tableau 2. Chemins les plus fréquentés par les phrasettes dans les partitions de 2 à 10 classes

Les phrasettes sont reliées par 295 chemins différents alors qu'il y a 10 ! possibilités (soit plus de 3 millions de chemins), ce qui montre qu'il y a des chemins privilégiés entre les différentes partitions. On peut remarquer que certaines classes n'apparaissent pas dans les 10 chemins les plus fréquentés : ce sont les classes 8 et 10. Aux dix premiers chemins sont associés 3 508 phrasettes, soit 75% du corpus.

Sur la figure 1 suivante, le treillis associé aux phrasettes est représenté sous forme de graphe. Cette figure montre comment se stabilisent les classes au fur et à mesure des partitions. On peut remarquer que la classe 1 est et reste une classe stable dès la partition en 4 classes. La classe 2 reste stable de la partition en 3 classes à la partition en 5 classes, elle se subdivise ensuite. La partition en 7 classes correspond à un niveau de stabilité pour toutes ses classes, car de cette partition il n'y a qu'une classe qui se partage pour créer la classe 8 dans la partition en 8 classes, les autres classes ne sont pas modifiées dans cette nouvelle partition.

4.2. *Extraction de l'arbre*

Visuellement, il est très facile de créer un arbre à partir des partitions. Il faut repérer des états stables et construire l'arbre à partir de ces états. Par exemple à partir du treillis précédent, si on s'arrête à la partition 8 on obtient l'arbre suivant, sur la figure 2. On part de la partition P8 de la figure 1 et on réalise des coupures dans l'arbre quand des branches s'isolent et des fusions quand une classe d'une partition est issue de plus d'une classe de la partition d'ordre immédiatement inférieure (voir détails dans la section suivante consacrée au critère de stabilité des classes). La fusion des classes 4, 5, 7 et 3 de P7 correspond à la fusion des classes 4, 5 et 3 des partitions P6 et P5 et à la fusion des classes 4 et 3 de P4, on leur associe donc un seul nœud dans la nouvelle arborescence qui sera étiqueté par les étiquettes des classes au niveau le plus proche de la racine soit ici : « 4,3 (P4) ». On fusionne dans cette branche les classes 1 et 3 de P3 dans une seule classe avec l'étiquette « 1,3 (P3) ». Les classes 6 et 8 de la partition P8 sont fusionnées dans la classe 6 des partitions P7 et P6, l'étiquette, établie selon la même règle que précédemment est « 6 (P6) ». Les classes 2 et 6 des partitions P7 et P6 fusionnent dans la classe 2 de la partition P5 que l'on retrouve identique dans P4 et P3, donc dans l'arbre avec l'étiquette « 2 (P3) ». La classe 1 de la partition 8 fusionne avec la classe 3 de la partition P3, elle porte l'étiquette « 1 (P4) » car c'est à partir de la partition P4 que cette classe s'isole définitivement.

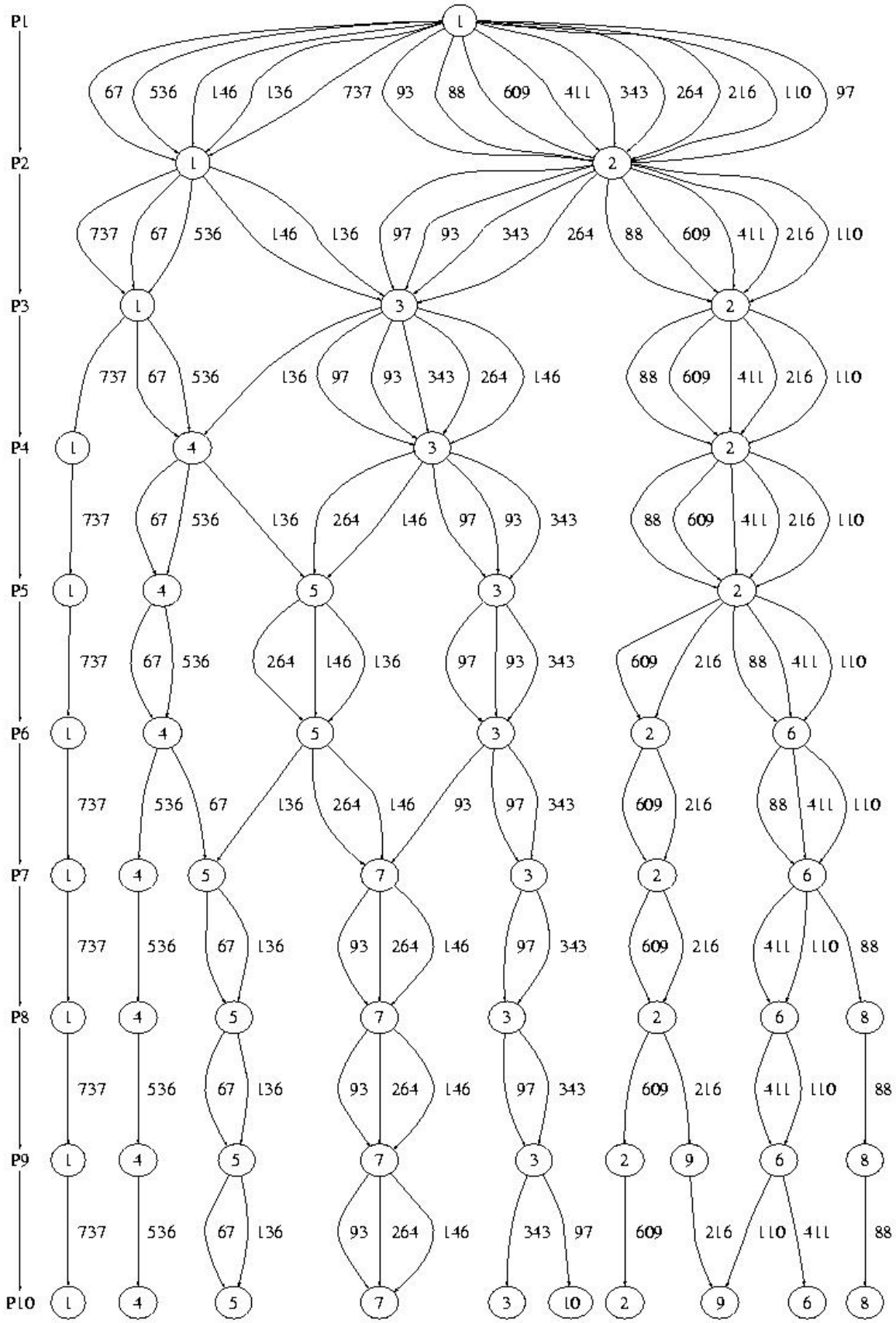


Figure 1. Treillis formé par les chemins des phrasettes dans les partitions de 2 à 10 classes

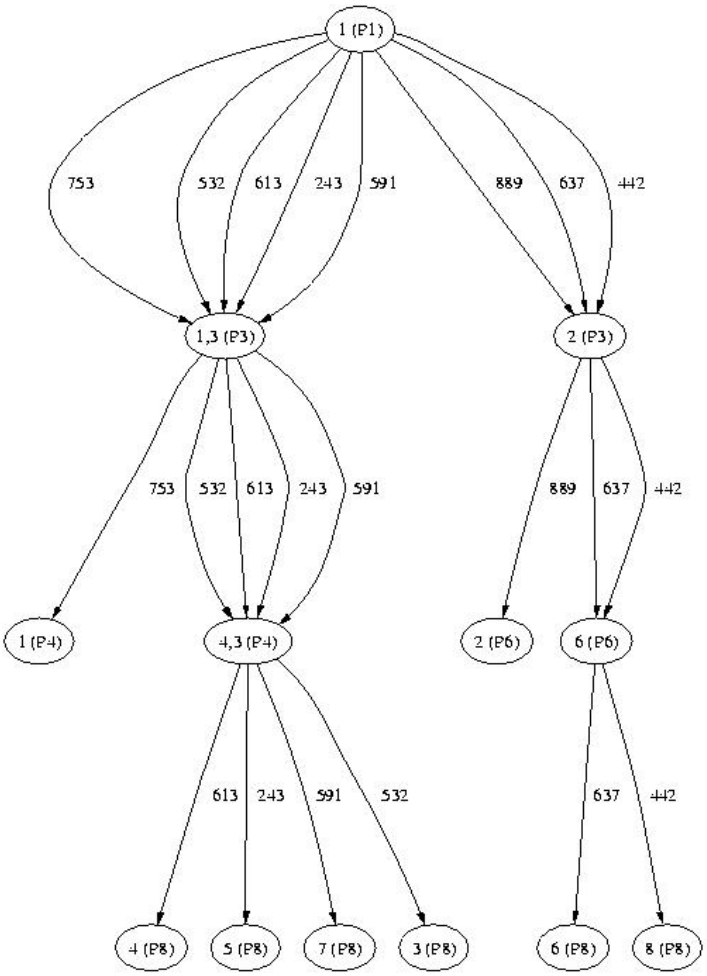


Figure 2.
Arbre extrait du treillis des chemins des phrasettes dans des partitions de 2 à 8 classes. Sur chaque branche figure le nombre de phrasettes qui transitent par les classes qui sont les nœuds de ce graphe

4.3. *Critères de stabilité des classes*

Des critères automatiques peuvent être utilisés pour trouver les classes stables dans les partitions successives. Nous en mettons deux en évidence, l'un fondé sur les chemins entre partitions, l'autre sur l'entropie des classes.

4.3.1. *Critère 1 : chemins entre deux partitions successives*

Une classe d'une partition en k classes sera dite stable si tous ses éléments (textes) sont issus d'une même classe de la partition en k-1 éléments, en d'autres termes, s'il n'existe qu'un chemin pour accéder à la partition k en venant de la partition k-1. Si toutes les classes de la partition k sont stables il y a k chemins entre les partitions k-1 et k, par contre si toutes les classes sont instables, il y a (k-1)*k chemins entre les partitions k-1 et k. Ainsi le nombre de chemins observés entre deux partitions successives peut être considéré comme un indicateur de stabilité des classes : pour une partition en k classes, il varie entre k (stabilité totale) et k*k-1 (instabilité totale), cet observable est un indice pour choisir une valeur de K.

Dans la phase exploratoire, il est nécessaire de choisir un nombre K assez grand pour observer s'il y a des zones de stabilité dans le nombre de chemins observés. A chaque texte on

associe la liste des classes auxquelles il appartient dans les partitions successives. On obtient ainsi tous les chemins observés que l'on compte ensuite. La courbe suivante, figure 3, montre l'évolution du nombre de chemins pour des partitions successives des phrasettes pour k variant de 1 à 15.

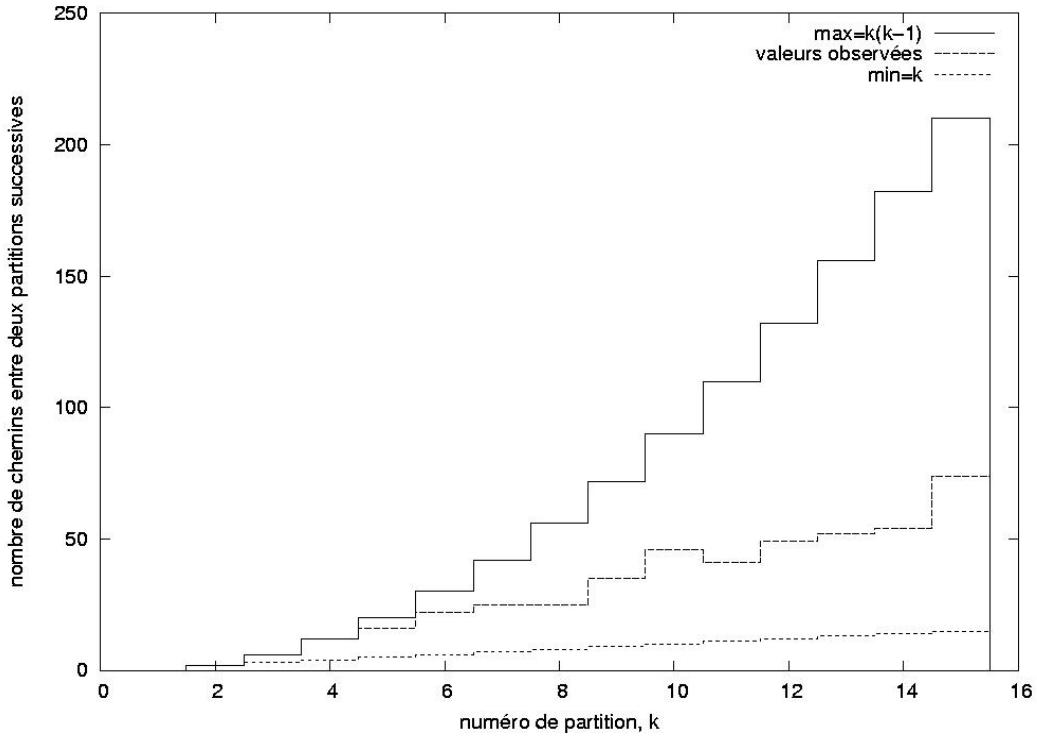


Figure 3. Nombres de chemins observés entre les partitions successives de phrasettes, comparés aux valeurs théoriques minimum et maximum

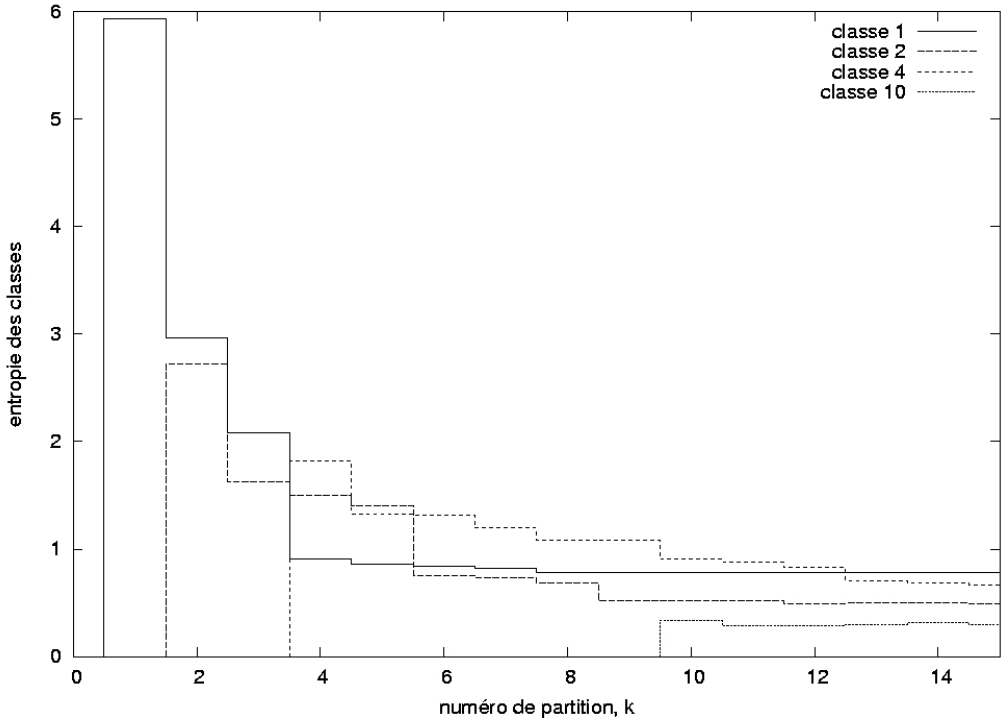


Figure 4. Variations d'entropie des classes 1, 2, 4 et 10 dans les partitions successives

On remarque qu'il y a stagnation du nombre de chemins entre les partitions P7 et P8 et une diminution de ce nombre entre P10 et P11, ce qui indique une stabilisation globale. La stagnation entre P7 et P8 correspond à la phase de stabilisation déjà observée dans le treillis de la figure 1.

4.3.2. *Critère 2 : entropie des classes*

On peut observer plus finement les transitions pour chaque classe, par exemple, comment évolue la classe 1 au travers des partitions, la classe 2 … Sur la figure 4 suivante nous montrons trois variations typiques de l'entropie dans les différentes partitions des phrasettes.

Pour la classe 1, l'entropie varie très rapidement en 3 étapes pour rester ensuite quasiment stable jusqu'à la partition en 15 classes.

Pour la classe 2, l'entropie varie fortement de la partition en 2 classes à la partition en 3 classes, elle varie peu entre les partitions de 3 à 5, puis de nouveau il y a une variation importante de P5 à P6, ensuite les variations sont faibles. On peut en déduire que la classe 2 est relativement stable dans les partitions 3, 4 et 5, que dans la partition 6, elle se divise pour redevenir stable dans les partitions 7 et 8. Elle se subdivise de nouveau à la partition 9 puis reste stable. La classe 2 semble être une classe homogène qui se subdivise.

L'entropie de la classe 4 chute fortement dans la partition P5, varie peu de P5 à P6 puis chute de nouveau avec P7. Les deux « chutes » correspondent à des divisions visibles sur le graphe de la figure 1 : division de la classe 4 (P4) en classes 4 (P5) et 5 (P6), de la classe 4 (P6) en classes 4 (P7) et 5 (P7).

Les variations d'entropie de la classe 10 montrent que des accroissements de l'entropie sont possibles au niveau local d'une classe.

On observe donc une variation de l'entropie fortement corrélée au parcours du graphe : de grandes variations correspondent à des divisions, des stagnations correspondent à des classes stables ou à des regroupements. Il est alors difficile avec ce critère de séparer ces deux derniers phénomènes. C'est d'ailleurs pour cela que la variation du critère d'optimisation qui est généralement utilisée (Duda et Hart, 1973) pour fixer le nombre de classes ne donne pas complètement satisfaction. Il faut plutôt le considérer comme un critère d'appoint par rapport au critère précédent.

5. Conclusion

Nous avons montré comment extraire une structure thématique latente dans des données : à partir de partitions successives de textes, de 2 à K classes, des classes stables dans ces partitions sont détectées et servent alors de point d'ancrage pour constituer un arbre éventuellement non équilibré avec des ramifications variées qui reflète la structure latente des données. On cherche d'abord un ordre de partitionnement (entre 2 et K) dans lequel les classes sont stables. A partir du treillis généré par les chemins des textes dans ces partitions successives, on génère un arbre par fusions et troncatures selon les branchements entre les partitions successives. Cette méthode originale peut être adaptée pour rechercher des classes stables à des niveaux de granularité plus fins, on peut utiliser d'autres fourchettes de variation, de K à K', et incrémenter ce nombre de classes par pas plus grand que 1. Elle doit également être complétée pour que la découverte de la structure soit entièrement automatisée à partir des critères de stabilité des classes décrits ci-dessus. Cette méthode non supervisée est utilisable sur tout type de données, par exemple, on peut envisager de créer une arborescence de mots à partir de profils de mots dans des textes ou de structurer des ensembles de pages Web.

Cette étude a été réalisée dans le cadre d'un projet BQR de l'Université Paris XI (Orsay, France) et a donné lieu à une première évaluation. Un fichier XML représentatif de la structure trouvée a été généré. A l'aide du formalisme de requêtes XPath, quelques questions ont été formulées pour tester l'apport de la hiérarchisation des informations touristiques. Dans tous les cas, la hiérarchisation améliore la précision de la recherche. Par exemple, pour la recherche de « séjour au bord d'un lagon », 19 fiches ont été trouvées dans la base de données XML pour 48 fiches retournées à partir des données non structurées. Les fiches supplémentaires correspondent à des « piscines à lagon » !

Références

Celeux G., Diday E., Govaert G., Lechevallier Y. et Ralambondrainy H. (1989). *Classification automatique des données.* Dunod.

Cover T. et Thomas J. (1991). *Elements of Information Theory*. Wiley & sons.

Duda R.O. et Hart P.E. (1973). *Pattern classification and Scene Analysis*. Wiley & sons.

Jardino M. et Adda G. (1993). Automatic word classification using simulated annealing. In *Proceedings of ICASSP'93*, Minneapolis, USA: 1191.

Jardino M. (2000). Unsupervised non-hierarchical entropy-based clustering. In Kiers H.A.L., Rasson J.-P., Groenen P.J.F. et Schader M. (Eds), *Data Analysis, Classification and Related Methods.* Springer : 29.

Jelinek F. (1988). *Statistical Methods for Speech recognition*. MIT Press.

Kaufman L. et Rousseeuw P.J. (1990). *Finding groups in data*.Wiley & sons.

Lebart L., Morineau A. et Piron M. (1995). *Statistique exploratoire multidimensionnelle.* Dunod.

Saporta G. (1990). *Probabilités, analyse des Données et Statistique*. Technip.

Analyse grammatico-métrique d'une monographie « multi-générique » ; le substantif

Margareta Kastberg Sjöblom

ILF-CNRS, Bases, Corpus et Langage (UMR 6039)
UFR Lettres, Arts et Sciences humaines, 98, Bd Edouard Herriot, 06204 NICE Cedex 3
kastberg@unice.fr

Abstract

The purpose of this paper is to explore an analysis of the parts-of-speech of a literary corpus. Today's techniques allow, by the integration of lemmatizers into the statistical tool, direct access to normalized and tagged textual data, as well as to the indispensable grammatical codes. This automatic sorting of the grammatical divisions paves the way to an impartial overview of the specific and personal style of an author as well as to the genre-related variations. The grammatical distribution, by its unconscious character, seems to show more subtle nuances than the traditional thematically orientated study and provides new valuable elements to stylistic studies.

Résumé

Cet article s'intéresse à l'étude des parties du discours d'un corpus littéraire contenant des textes de genres différents. Désormais l'intégration de lemmatiseurs dans les logiciels lexicométriques permet, grâce à l'accès aux codes grammaticaux, le recensement automatique et impartial des divisions grammaticales, personnelles et caractéristiques des typologies de textes diverses. Cette distribution, qui est bien un critère de distinction des œuvres, manifeste peut-être des choix plus subtils que celui du vocabulaire – en tout cas moins liés à la thématique de chaque ouvrage, ce qui peut apporter à l'analyse des éléments nouveaux.

Mots-clés : lexicométrie, lemmatisation, stylométrie, stylistique de corpus, genres littéraires, typologie textuelle, Le Clézio.

1. Introduction

Les parties du discours, ou les catégories grammaticales, sont une classification des mots qui nous a été transmise par la syntaxe traditionnelle, basée sur trois critères qui sont la nature, la fonction et la position. Quel est l'intérêt d'analyser la distribution des parties du discours dans un corpus littéraire « multi-générique » informatisé, et de quelle façon cette analyse peut-elle contribuer à la compréhension de l'utilisation et de la fonction des catégories grammaticales ?

« Investigating this question, écrit Douglas Biber (1998 : 57), can help to understand how different varieties exploit the grammatical categories of words available to them… Using the corpus, we can analyze the distribution and function of these different categories of words and study the part that they play in fulfilling the communicative function of different registers."

En effet, la distribution des parties du discours n'est pas constante, elle varie selon les époques, les auteurs et les genres. Déjà l'étude de Guiraud (1954) montre que la distribution des catégories grammaticales dans les ouvrages littéraires est fortement influencée par le genre et l'époque où ils s'inscrivent. Cette distribution, qui est bien un critère de distinction des œuvres, manifeste peut-être des choix plus subtils que celle du vocabulaire – en tout cas moins liés à la thématique de chaque ouvrage, ce qui peut apporter à l'analyse des éléments très révélateurs. Peut-on à partir de l'étude des parties du discours mieux comprendre l'écriture et

les différentes typologies de textes ? Quels sont les particularités grammaticales qui contribuent au style d'un écrivain et quel est le rôle du genre littéraire dans leur distribution ? Dans cette étude qui s'appuie sur une monographie contemporaine « multi-générique » – l'œuvre de J.M.G. Le Clézio (annexe 1) – nous nous attarderons, après avoir étudié la distribution des catégories grammaticales, sur quelques aspects morphologiques et syntaxiques qui nous ont paru intéressants et révélateurs dans la catégorie nominale.

Malheureusement les parties du discours ne sont pas des catégories entièrement stables et étanches les unes aux autres. Si dans chaque contexte l'appartenance d'un mot à une catégorie précise s'impose sans grande difficulté, en revanche le même mot rentrera dans une autre catégorie pour peu que change le contexte, même si le contenu sémantique reste constant. Les cas de *Venir pour dîner* et *Venir pour le dîner* ou bien *Tout est bon* et *Le tout est bon* sont des exemples de ce phénomène relativement courant dans la langue française. En effet, les fondements et les méthodes sur lesquels s'appuient les études des catégories grammaticales varient d'un ouvrage à l'autre. Il convient donc, avant de se lancer dans l'étude de catégories grammaticales de s'attarder un peu sur les principes et les techniques qui guident cette étude.

2. Les méthodes d'analyse

Désormais la quantification et la lemmatisation des corpus ouvrent la voie à cette composante essentielle de l'écriture qu'est l'étude de la distribution des parties du discours. Cette analyse, qui demande l'accès à la forme canonique du mot, au lemme, ne peut guère se fonder sur la distribution des effectifs d'un corpus s'appuyant sur la forme graphique. C'est la lemmatisation qui permet d'étiqueter le corpus selon les catégories grammaticales et de classer les éléments du vocabulaire selon leur appartenance à une catégorie spécifique. Les codes grammaticaux fournis par l'étiqueteur morphosyntaxique au cours de l'opération de lemmatisation « automatique » constituent ici un outil indispensable.

Toutefois, la lemmatisation automatique repose sur des critères préétablis selon des méthodes d'analyse différentes et les résultats peuvent varier selon le choix de l'étiqueteur et de ses principes. Il faut également savoir que la lemmatisation automatique n'est pas sans erreurs : erreurs d'analyse contextuelles, mauvaise définition de certaines classes grammaticales, etc. Il est donc recommandé de tenir compte de ces imperfections lors de l'exploitation des résultats. Les parties du discours distinguées par les logiciels d'analyse morpho-syntaxique sont toutefois nettement plus fiables que les fonctions grammaticales car leur reconnaissance est plus facile. Pour beaucoup de mots qui ne sont pas sujets à l'homographie, le codage est automatique et indépendant du contexte : la proposition du dictionnaire, unique, est immédiatement acceptée. Et quand il s'agit de diviser les homographes en catégories différentes (par exemple les cas très nombreux, du type *la marche/il marche*, où le substantif peut se confondre avec un verbe), une analyse syntaxique simple, désormais quasi-automatique (selon les différents logiciels), emporte la décision. D'autres cas peuvent être plus difficiles, comme la reconnaissance de l'adverbe *fort* dans le syntagme *un homme fort bon*. Face à ces ambiguïtés d'analyse, il y a différents types de lemmatiseurs : ceux qui prennent une décision dans tous les cas de figure (Cordial) et ceux qui rendent la main au linguiste (Logiciel Labbé). Le choix d'un type de lemmatiseur ou l'autre se fait en fonction de la taille de corpus (s'il est grand l'analyseur automatique offre plus d'avantages que d'inconvénients, s'il est de taille plus modeste l'autre solution est peut-être préférable). Nous avons pratiqué les deux solutions dans notre thèse, ce qui nous a permis de constater d'ailleurs que les résultats restaient assez stables de l'un à l'autre. L'étiquetage morphologique est généralement plus fiable que le « parsage » syntaxique qui recourt à des modèles et des procédures complexes.

Dans cette étude nous exploitons un corpus qui englobe la quasi-totalité de l'œuvre de Le Clézio (Cf. Kastberg Sjöblom 2002 : 100-110) lemmatisé avec l'analyseur Cordial 7, désormais intégré au logiciel Hyperbase, le logiciel lexicométrique sur lequel s'appuie cette analyse.

3. La distribution des catégories grammaticales du corpus

L'analyse morpho-syntaxique Cordial aboutit à quelque 200 codes grammaticaux différents en utilisant toutes les combinaisons possibles. Le logiciel Hyperbase regroupe par la suite de façon « automatique » les codes et fournit la liste des fréquences.

Nous en avons extrait les 11 catégories fondamentales parmi celles que propose le programme Cordial ; verbes, substantifs, adjectifs, déterminants, pronoms, numéraux, interjections, prépositions, adverbes, conjonctions et délimiteurs (signes de ponctuations). Cette extraction met en évidence la richesse en substantifs (20,5 %) de notre corpus, une des caractéristiques des corpus de langue française. Les catégories des déterminants, des signes de ponctuation et des verbes occupent chacune entre 14 et 15 % du corpus. Quant aux pronoms et aux prépositions, ils se regroupent autour de 10 % chacun tandis que les adverbes, les adjectifs et les conjonctions se rassemblent autour de 5 % chacun. Les catégories de numéraux et d'interjections sont dérisoires par leur nombre. Pierre Guiraud (1954 :104) avait déjà observé que dans la littérature le nombre de substantifs et celui des verbes varient en proportion inverse, le substantif étant dominant dans la prose abstraite et le verbe dans les récits. La richesse du substantif dans l'écriture lecléziennne, caractéristique partagée avec les grands écrivains du XIXème siècle, est étonnante car on aurait pu croire que les romans et les nouvelles de Le Clézio étaient plus proches des récits et donc devaient favoriser le verbe.

Si nous regardons d'un peu plus près les catégories grammaticales dans notre corpus, nous constatons que leur distribution n'est pas régulière. L'analyse factorielle de la liste de fréquences des différentes codes grammaticaux illustre des variations importantes :

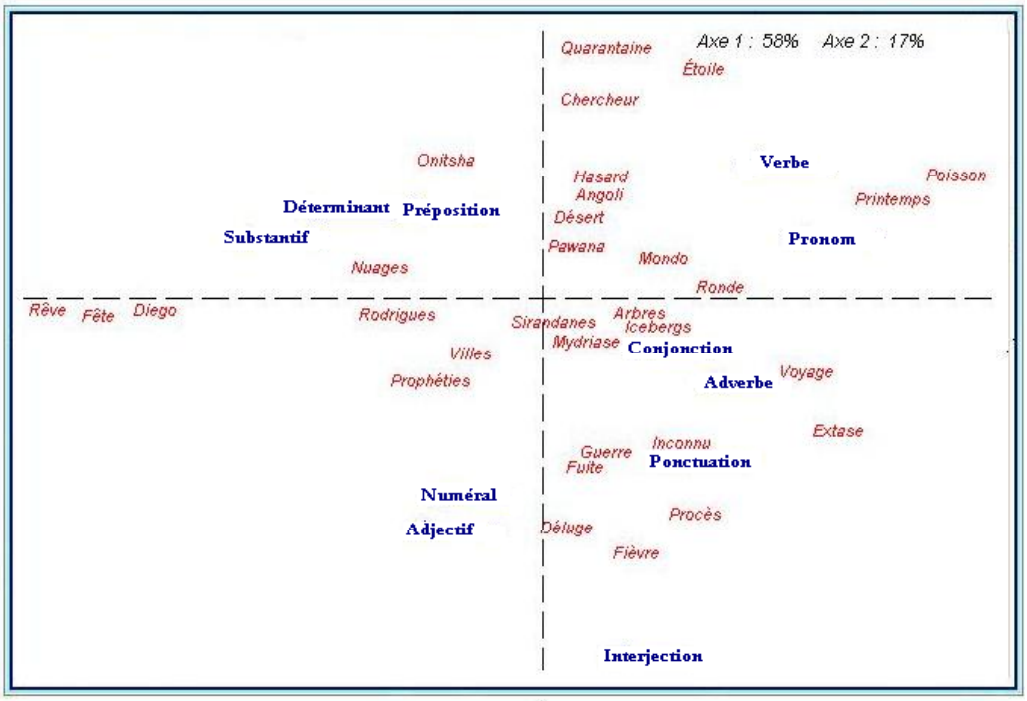


Figure 1.
Analyse factorielle de la distribution grammaticale du corpus selon la lemmatisation par Cordial 7.

Le premier facteur oppose la catégorie verbale à la catégorie nominale. Le substantif à gauche attire les prépositions, les déterminants et les adjectifs tandis que le verbe en haut à droite attire les pronoms et les adverbes. Le second facteur parcourt la chronologie de l'écrivain du bas vers le haut du graphique. Les premiers ouvrages se trouvent en bas du graphique autour des catégories secondaires qui témoignent d'une écriture foisonnante (adjectifs, adverbes et interjections). Les derniers romans se situent en haut du tableau autour des catégories fondamentales, témoignant peut-être d'un assagissement de l'écriture, d'un travail de simplification de style. L'analyse factorielle rend également compte de l'importante opposition générique. Les ouvrages ethnologiques se regroupent à l'extrême gauche du graphique, les premiers romans appartenant à l'école du « nouveau roman » en bas à droite, tandis que les œuvres fictionnelles se trouvent au centre supérieur du tableau. Les ouvrages qui se trouvent au milieu sont les plus courts, tous genres confondus. Examinons pour l'instant le phénomène dont témoigne le premier facteur de l'analyse factorielle, l'opposition du groupe verbal et du groupe nominal.

4. L'opposition des catégories grammaticales

On observe souvent dans un corpus clos, comme nous venons de le faire, que deux camps, la catégorie nominale et la catégorie verbale, s'affrontent : la classe du verbe et les catégories qui lui sont proches (subordonnants, relatifs, pronoms et adverbes) s'opposent à la classe nominale qui réunit autour du substantif les adjectifs, les déterminants, les prépositions et souvent les coordinations.

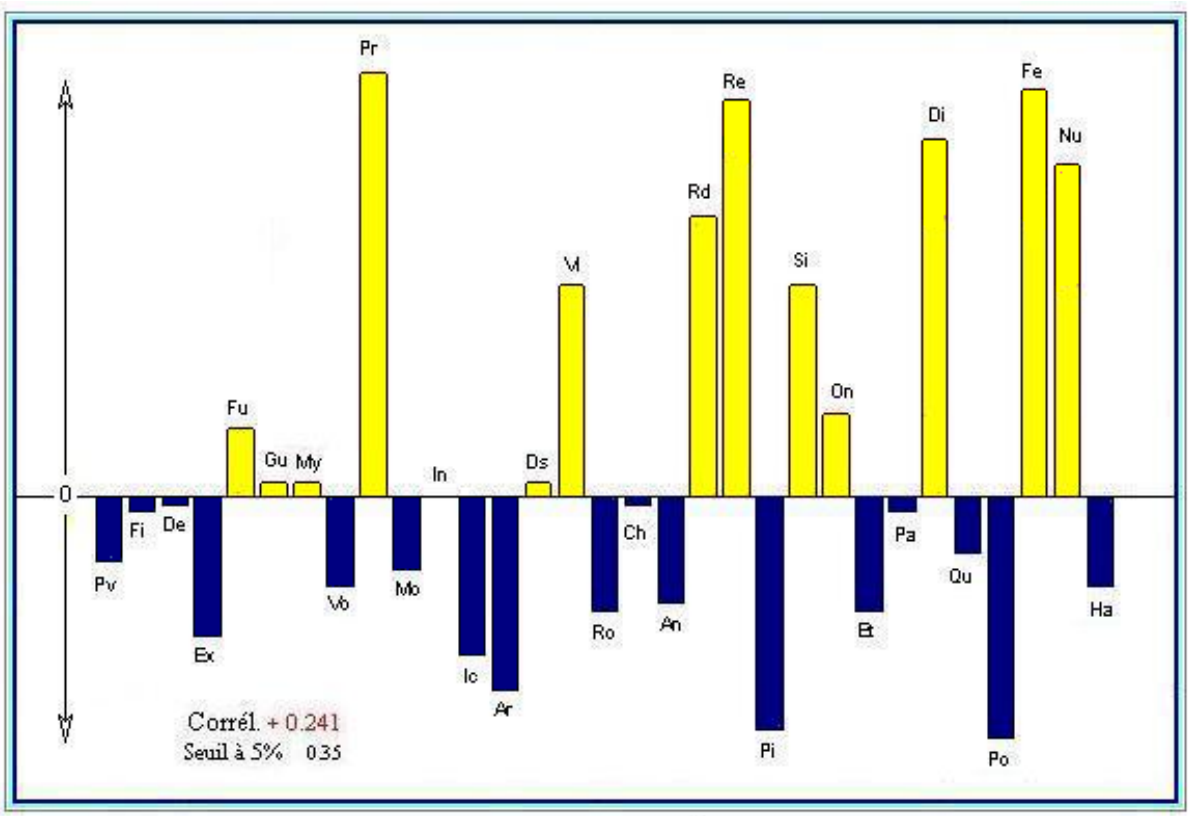


Figure 2. Histogramme du quotient substantifs/verbes.

L'histogramme du quotient entre les 459.957 substantifs et les 321108 verbes permet de voir de près ce que l'analyse factorielle nous a déjà indiqué : lorsqu'une œuvre est riche en subs-

tantifs, elle est pauvre en verbes et vice-versa. Il nous apporte la preuve formelle de l'opposition des deux catégories qui semble s'accentuer, avec des écarts de plus en plus importants et des mouvements de plus en plus amples, au fur et à mesure que l'œuvre progresse et des écarts qui se révèlent très sensibles au genre littéraire[1].

Au début de la production de l'écrivain, dans sa période « nouveau roman », les deux courbes ne s'écartent point, elles se suivent au contraire, les deux catégories étant déficitaires dans cette partie de l'œuvre. C'est à partir de l'essai *L'extase matérielle* que l'opposition se déclare. Les écarts les plus importants – avec un déficit important de verbes et un grand excédent de substantifs – sont à trouver dans les ouvrages d'ethnologie et dans les essais qui traitent du nouveau monde, comme *Le rêve mexicain* ainsi que dans la biographie *Diego et Frida*. *Poisson d'or* est le seul roman de cette époque qui présente un écart d'une grande amplitude, mais l'écart cette fois-ci témoigne d'un déficit important de substantifs et d'un excès de verbes.

Dans les œuvres non fictionnelles – les ouvrages ethnologiques, les essais, les récits de voyage et la biographie – l'évolution de l'opposition entre la catégorie du substantif et celle des verbes est en effet assez spectaculaire. Au début, les substantifs sont déficitaires et les verbes excédentaires, mais assez vite les rôles s'inversent et l'écart s'amplifie de façon importante. Il est difficile de fournir une explication précise, mais à un moment qui correspond à la découverte de la culture amérindienne et mexicaine, capitale pour notre écrivain, les substantifs commencent à abonder tandis que les verbes diminuent de façon considérable. Peut-être, à partir de ce moment, n'y a-t-il plus besoin du mouvement, des dialogues ni des verbes (d'action ou de parole) ; il suffit de regarder et Le Clézio observe, décrit et partage ce qu'il voit avec ses lecteurs en recourant à de nombreux substantifs.

5. Le syntagme nominal

Nous avons déjà pu constater l'ampleur de la catégorie nominale dans notre corpus, un phénomène qui est constaté dans pratiquement toutes les études statistiques sur le vocabulaire français[2]. L'histogramme ci-dessous rend compte de sa distribution relative à travers l'œuvre :

Nous pouvons observer que la fréquence des substantifs à l'intérieur du corpus n'est pas constante. Le graphique rend compte de deux dynamiques différentes : la différenciation générique et l'évolution dans le temps. Concernant l'opposition des genres littéraires, la tendance générale est celle d'un substantif déficitaire dans les romans tandis qu'il est excédentaire dans les ouvrages d'ethnologie, dans la biographie et dans les récits de voyages. En revanche, dans les livres qui favorisent le dialogue, et par conséquent une langue plus orale, le substantif est déficitaire. Des livres comme *Printemps et autres saisons* et *Poisson d'or* ne sont pas seulement riches en dialogues mais sont également des livres où l'action prime sur la description.

[1] Le quotient est le rapport entre les deux séries. Il permet de voir comment se séparent les parallèles quand deux séries sont liées et parallèles. Comme les deux séries peuvent avoir un poids très inégal, la seconde est d'abord ramenée à la dimension de la première, proportionnellement, pour que le total des deux séries soit le même. Le quotient est calculé ensuite terme à terme, et s'équilibre nécessairement autour de la valeur 1.

[2] Selon Marc Hug (1989 : 5) : « Le syntagme nominal est l'unité qui s'impose d'emblée quand on commence l'étude statistique : non seulement c'est l'unité la plus abondamment représentée dans tout énoncé, et une unité très multiforme, mais c'est en même temps une unité fragmentaire, qui peut être étudiée indépendamment de la structure d'ensemble de la phrase… »

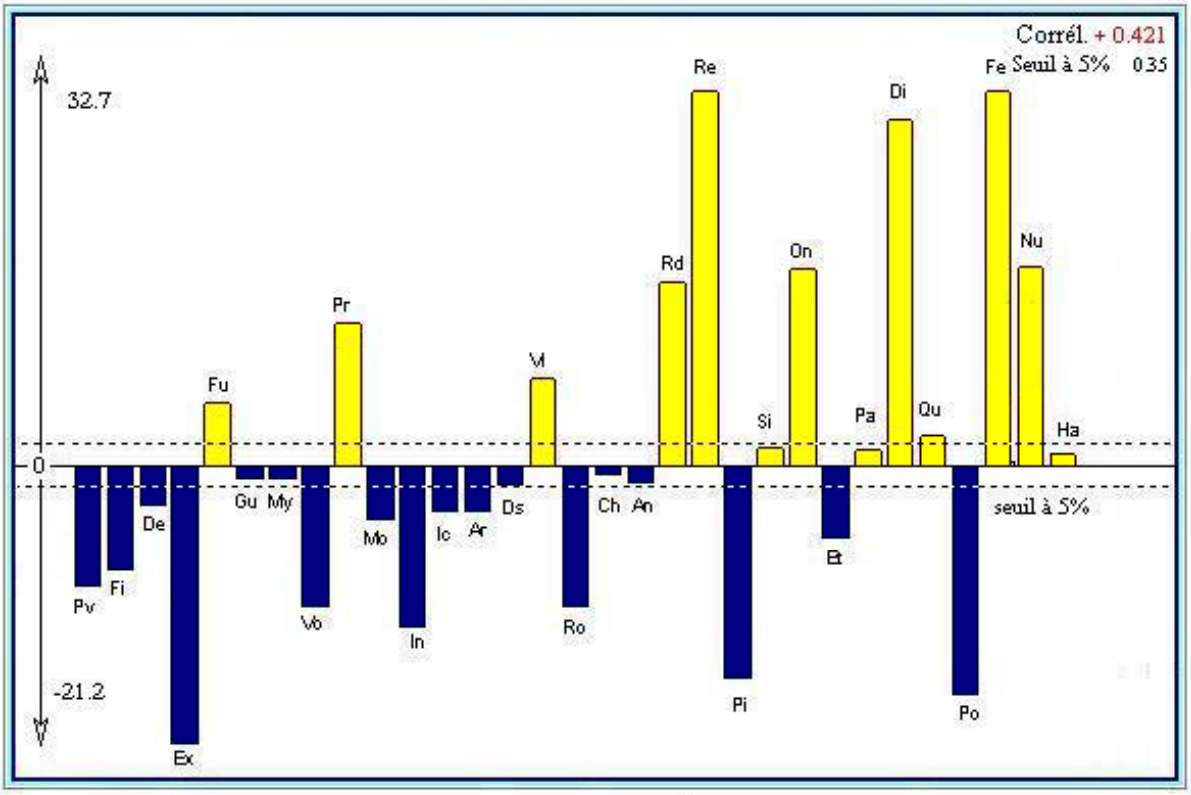


Figure 3. La distribution des substantifs dans le corpus(écarts réduits).

Ces résultats sont confirmés par les constats qu'a faits Étienne Brunet à propos du vocabulaire français de 1789 à nos jours (1981 : 304) concernant « le style substantif ». Le langage technique et le discours intellectuel (que l'on trouve dans les essais et dans les ouvrages ethnologiques de Le Clézio) utilisent plus volontiers les substantifs car ils sont plus adaptés aux définitions abstraites et « s'enchaînent avec une grande rigueur logique dans des phrases reflétant sous une forme très ramassée une pensée complexe. »

Il est intéressant de comparer cette analyse avec celle de la longueur de la phrase que nous avons effectuée sur ce même corpus (Cf. Kastberg Sjöblom, 2002a : 230-240). Les deux histogrammes sont pratiquement superposables : celui de la distribution des substantifs et celui de la longueur de la phrase. Cela confirme ce que nous avons pu constater auparavant, c'est-à-dire que la construction de la phrase longue chez Le Clézio repose souvent sur l'énumération et l'accumulation, et ce sont, bien évidemment, des substantifs qui prolifèrent dans ce système. Nous notons également chez Le Clézio la même corrélation entre la fréquence du substantif et l'accroissement lexical ainsi que la richesse du vocabulaire qu'a déjà trouvée Charles Muller (1967 : 111-115) dans ses études sur le vocabulaire de Corneille, où les pièces les plus riches en substantifs avaient en général aussi un vocabulaire étendu, bien qu'il y eût des exceptions à cette règle.

La division de la catégorie de substantifs en genres, c'est-à-dire en masculin et en féminin, pourrait nous apporter plus de précisions sur l'usage du nom dans l'œuvre de Le Clézio[3].

[3] Dans les études statistiques s'appuyant sur des corpus non lemmatisés, l'analyse du genre masculin et féminin se faisait surtout à partir de l'étude des articles, c'est-à-dire *la*, *le*, ou *une*, les formes *l'* et *les* ne disant rien sur le genre. Or ces formes sont ambiguës et des confusions avec les numéraux aussi bien qu'avec des pronoms personnels rendaient la statistique moins fiable. Les formes contractées au et du étaient délicates à traiter et le phénomène de la neutralisation de l'opposition du genre masculin-féminin dans les formes *l'*, *les*, *aux* et *des* ne facilitait certes pas la tâche.

Désormais, le logiciel Hyperbase dans sa version lemmatisée permet d'extraire les noms féminins et les noms masculins distinctement grâce à l'analyseur incorporé. Si nous divisons la catégorie des substantifs en genres, nous pouvons constater que le masculin domine avec 56,5 % des effectifs, le féminin n'en comportant que 43,5 %. L'analyse de la distribution relative permet de constater que la distribution des deux genres est relativement parallèle et qu'il n'y a pas d'opposition franche entre le masculin et le féminin (coefficient de corrélation significativement positif (+0.60)).

À la comparaison avec les valeurs obtenues lors de l'analyse des deux articles *le* et *la*[4], nous pouvons constater que l'article *le* emporte sur l'article *la* avec 57 % contre 43 %, c'est-à-dire un résultat quasi identique à celui du masculin et du féminin. Les courbes obtenues font preuve d'un parallélisme encore plus fidèle que celles des substantifs, avec un coefficient de corrélation de +0,749. Dans des études antérieures on avait presque toujours pu observer que l'article féminin l'emportait sur le masculin avec des proportions autour de 55 % contre 45 % (Brunet, 1988 ; Engwall, 1982)[5]. La préférence pour l'article féminin est encore plus marquée dans *Frantext* avec 56 % contre 44 % pour le masculin. La prépondérance de l'article *le* par rapport à *la* semble vraiment être une caractéristique de notre écrivain.

Comment interpréter ces résultats ? Nous savons qu'en français l'opposition des genres n'a pas de motivation précise. Toutefois dans certains substantifs on peut observer une régularité en genre due à la présence de suffixes. Les suffixes *-tion*, *-té* ainsi que *-ade* marquant une action ou une qualité imposent le féminin, tandis que *-ment* et *-isme* exigent l'article masculin. Étienne Brunet écrit ceci à propos de la prépondérance habituelle de l'article féminin (1988 : 193) : « L'abstraction intéresse plus l'article défini que l'indéfini (on dit moins souvent une justice que la justice) et concerne davantage le féminin que le masculin. […] Le rapport *la*/*le* serait donc une mesure indirecte du degré d'abstraction d'un texte ou d'un corpus. »

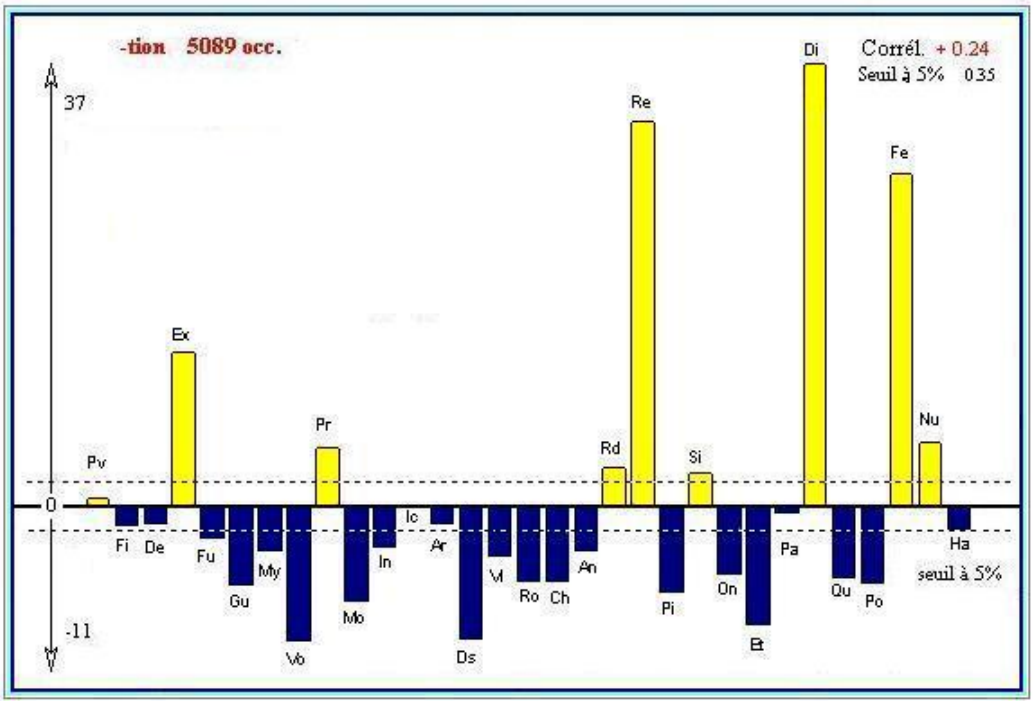


Figure 4. La distribution relative du suffixe -tion *(écarts réduits).*

[4] La forme unique du pluriel est exclue de l'analyse, ne permettant pas la distinction du genre en français.

[5] Rappelons qu'il s'agit dans son étude de formes graphiques et qu'il faut en tenir compte lors de la comparaison.

Les valeurs excédentaires du suffixe *-tion* dans notre corpus se trouvent dans les ouvrages ethnologiques et dans les essais où effectivement plus de place semble être donnée à l'abstraction et au discours intellectuel.

Dire à partir de ces analyses que Le Clézio fuit l'abstraction dans son œuvre romanesque serait sans doute une surinterprétation. Mais il y a chez notre auteur une volonté manifeste de rester dans le concret et de garder un langage simple et « terre à terre » ; il y revient souvent dans les divers entretiens et même dans ses textes et l'analyse statistique le confirme. Toutefois, l'histogramme de l'article *la* seul témoigne d'une tendance croissante vers l'abstraction, avec un coefficient de corrélation significativement positif :

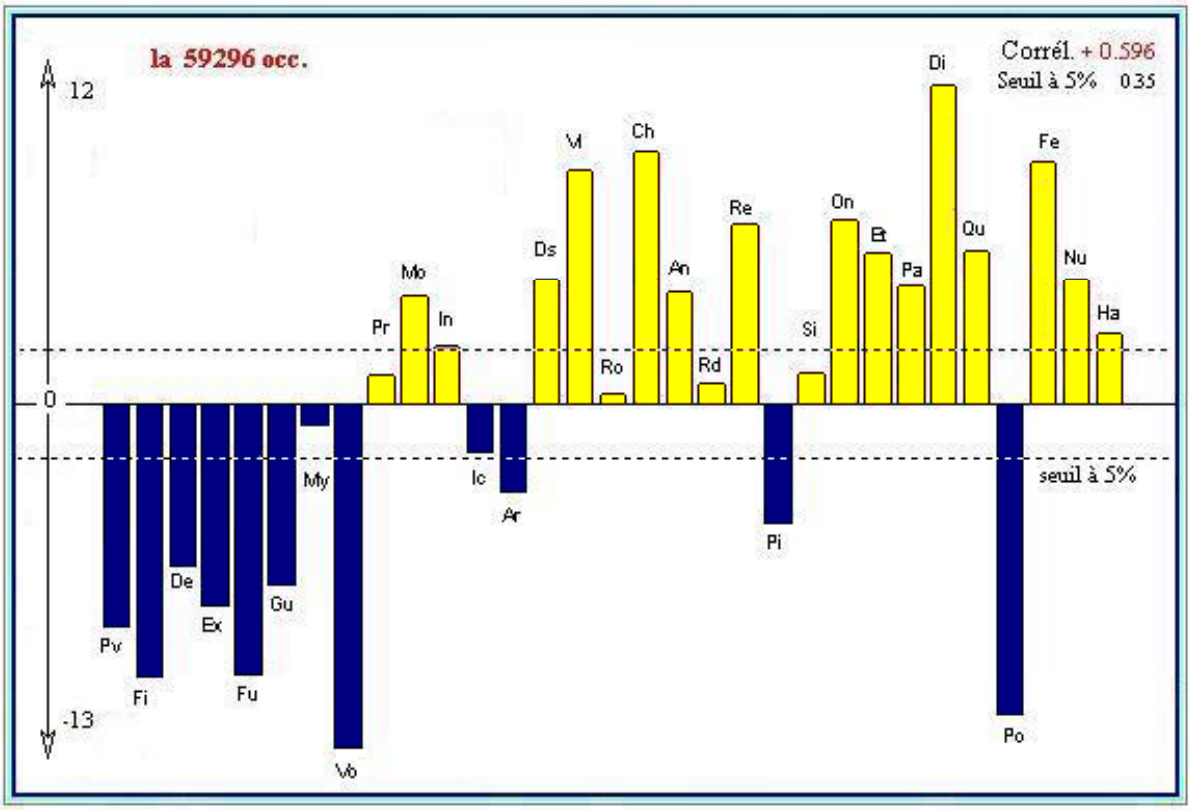


Figure 5. La distribution relative de l'article la dans le corpus (écarts réduits).

Que la première période « nouveau roman » soit largement déficitaire n'est pas étonnant avec cette écriture particulière qui ne favorise pas l'abstraction. Bien au contraire, elle est favorable au concret, car c'est une réalité bien concrète, la société de consommation avec ses attributs et l'envahissement d'objets, que Le Clézio dénonce durant cette période. Par la suite, l'écriture semble tendre de plus en plus vers l'abstraction dans tous les genres littéraires (accentuée dans les ouvrages ethnologiques, les essais et la biographie), à l'exception des livres riches en dialogues et en oralités comme *Voyage au pays des arbres*, *Printemps et autres saisons* et *Poisson d'or*.

6. Conclusion

Cette étude qui ne prétend qu'à donner quelques exemples d'utilisation des codes grammaticaux dans l'étude stylistique et typologique de textes et ne peut pas – faute de place – traiter tous les aspects sous lesquels on pourrait étudier les parties du discours et les éléments syntaxiques en exploitant cette technique. Mais elle aura montré qu'un excédent de l'une ou

l'autre catégorie dans un texte donné peut avoir plusieurs causes et produire de multiples effets stylistiques.

Malgré les difficultés rencontrées dues à l'opacité relative de l'analyseur Cordial, notamment à la façon dont il tranche entre le singulier et le pluriel du substantif et le fait qu'il faut admettre une certaine marge d'erreur ou d'incertitude sur les chiffres, ces doutes ne mettent pas en cause les grandes tendances qu'au contraire seule l'analyse statistique d'un grand corpus lemmatisé peut mettre en évidence.

En effet, le profil grammatical de J.M.G. Le Clézio qui émerge de nos différentes analyses est celui d'une écriture globalement marquée par une forte présence de la catégorie nominale mais qui change, qui évolue dans le temps et dont la richesse s'exprime dans plusieurs genres littéraires ; les variations génériques se reflètent en effet dans presque toutes les analyses que nous avons effectuées.

L'analyse lexicométrique et « grammatico-métrique » permet de mettre à jour ces procédés peu conscients que l'on a comparés aux empreintes digitales d'un écrivain. Les empreintes digitales pourtant ne changent pas avec le temps, au contraire, elles demeurent à jamais les mêmes. Notre étude montre que les procédés langagiers peuvent changer, même de façon significative, à l'intérieur d'une œuvre. Le jeune écrivain débutant ne fait pas forcement le même emploi des catégories grammaticales qu'il fera au sommet de sa production ou bien vers la fin de sa carrière et ce phénomène semble très accentué dans le cas de Le Clézio.

Une écriture qui change est une des caractéristiques fondamentales de notre corpus. En effet, il n'y a pas de « stabilisation » du style mais, au contraire, des écarts grandissants chez le Clézio. Dans l'œuvre de Le Clézio les procédés morphosyntaxiques ne sont pas statiques, les techniques d'expression évoluent et sont constamment mises en question.

La distinction de typologies de textes n'opère, nous semble-t-il, ni à un niveau conscient lors de la production, ni à un niveau interprétatif. L'opération de classification par laquelle un lecteur donne une certaine cohésion à une suite textuelle est une opération de lecture-interprétation qui confère au discours une certaine structure compositionnelle opèrant au niveau le plus profond, régi par la finalité des textes. Chaque genre littéraire a en fait son anatomie, sa physiologie et son fonctionnement au niveau pour ainsi dire « atomique » et cela transparaît très clairement dans les différents textes qui forment l'œuvre leclézienne.

Références

Biber D., Conrad S. et Reppen R. (1998). *Corpus linguistics, Investigating Language, Structure and Use*. Cambridge Approaches to Linguistics.

Brunet Ét. (1981). *Le vocabulaire français de 1789 à nos jours*. Champion – Slatkine.

Brunet Ét. (1988). *Le vocabulaire de Victor Hugo*. Champion-Slatkine.

Engwall G. (1982). Le vocabulaire dans les best-sellers des années 1960. In *Actes du 2° colloque de lexicologie politique*, Saint-Cloud, 15-20 septembre 1980, vol (1). Klincksieck.

Guiraud P. (1954). *Les caractères statistiques du vocabulaire : essai de méthodologie*. PUF.

Hug M. (1989). *Structures du syntagme nominal français, Étude statistique*. Champion-Slatkine.

Kastberg Sjöblom M. (2002a). *L'écriture de J.M.G. Le Clézio, une approche lexicométrique*. Université de Nice – Sophia Antipolis.

Kastberg Sjöblom M. (2002b). Le choix de la lemmatisation. Différentes méthodes appliquées à un même corpus. In *Actes des JADT 2002 :* 391-402.

Kastberg Sjöblom M. (2003). Analyse lexicométrique de l'opposition générique dans une perspective endogène. In Williams G. (Ed.), *Actes des IIIèmes Journées de la linguistique de corpus*. Presses Universitaires de Rennes.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. CERAT. Cahier n° 7.

Malrieu D. et Rastier F. (2002). Genres et variations morphosyntaxiques. In Angel Martin Municio (Ed.), *Actas del segundo seminario de la escuela interlatina de altos estudios en lingüística aplicada, Matemáticas y tratamiento de corpus, San Millán de la Cogolla, 19-23 septiembre de 2000*, Logroňo, Fundación San Millán de la Cogolla.

Muller Ch. (1967). *Le vocabulaire du théâtre de Pierre Corneille, Étude de statistique lexicale*. (réédition Slatkine 1993).

Picard J., Pibarot A. et Labbé D. (1995). Un outil de statistique textuelle : le lemmatiseur. In *Travaux scientifiques C.R.S.S.A.*, vol. (16) : 395-396.

Annexe 1

Genres	Titre	No	Genres	Titre	No
nouveau roman	Le procès-verbal	1			
	La fièvre	2	**récits poétiques**	Mydriase	7
	Le déluge	3		Vers les icebergs	12
	Le livre des fuites	5			
	La guerre	6	**essais littéraires**	L'extase matérielle	4
	Voyages de l'autre côté	8		L'inconnu sur la terre	11
				Trois villes saintes	15
romans traditionnels	Désert	14		Le rêve mexicain	20
	Le chercheur d'or	17			
	Angoli Mala	18	**ouvrages ethnologiques**	Les prophéties de Chilam Balam	9
	Voyage à Rodrigues	19		Sirandanes	22
	Onitsha	23		La fête chantée	29
	Etoile errante	24			
	La quarantaine	27	**enfant et jeunesse**	Voyages au pays des arbres	13
	Poisson d'or	28		Pawana	25
	Hasard	31			
			récits de voyages	Gens des nuages	30
nouvelles	Mondo et autres histoires	10			
	La ronde et autres histoires	16	**biographies**	Diego et Frida	26
	Printemps et autres histoires	21			

Création d'un espace conceptuel par analyse de données contextuelles

Nicolas Kumps, Pascal Francq, Alain Delchambre

CAD/CAM – Université Libre de Bruxelles – Av. F. D. Roosevelt, 50 – CP 165/14 1050 Bruxelles – Belgique
{nkumps, pfrancq, adelch}@ulb.ac.be

Abstract

The computerized treatment of numerical information requires a system of representation. The objective of the method presented in this paper is the entirely automated generation of a valid and compact system of representation of the information contained in documents. One hopes to group words referring itself on the same subject, the same idea in order to use these "concepts" to model information rather than to use the representation of these concepts.
The algorithm proposed generates a vector space of the concepts by gathering dimensions of a vector space of the terms. This regrouping is carried out by comparison of the parameters related to the context of appearance of the terms in the treated documents. The reduction of dimension of the space used will allow a faster data processing and requiring a less important memory capacity. This type of improvement being typically that required within the framework of real system, tallies of the application for which the method was developed.

Résumé

Le traitement informatisé de l'information numérique requiert un système de représentation. L'objectif de la méthode présentée dans cet article est la génération entièrement automatisée d'un système de représentation, valide et compact, d'informations contenues dans des documents. On espère grouper des mots se rapportant à un même sujet, une même idée afin d'utiliser ces « concepts » pour modéliser l'information plutôt que d'utiliser la représentation de ces concepts.
L'algorithme proposé génère un espace vectoriel de concepts en regroupant les dimensions d'un espace vectoriel de termes. Ce regroupement s'effectue par comparaison des paramètres liés au contexte d'apparition des termes dans les documents traités. La réduction de dimension de l'espace utilisé permettra un traitement de l'information plus rapide et demandant un espace mémoire moins important. Ce type d'amélioration étant typiquement celle recherchée dans le cadre de système réel, cadre de l'application pour laquelle la méthode a été développée.

Mots-clés : espace conceptuel, association de mots, similarité, cooccurrence.

1. Introduction

Ce travail s'inscrit dans le cadre d'un projet de recherche documentaire : le projet GALILEI[1]. Chaque utilisateur du système est représenté sous la forme d'un profil. Il pourra se définir autant de profils qu'il possède de centres d'intérêt. Il pourra alors juger les documents, au moyen d'un simple clique de souris, comme pertinent ou hors sujet par rapport à un de ces profils. Ceux-ci sont ensuite groupés en communautés virtuelles sur base des documents jugés. Un échange de documents pertinents est alors possible au sein d'une même communauté. Le fruit de la recherche documentaire de tous est ainsi partagé entre les utilisateurs du système (Francq, 2003).

[1] Subventionné par la Région wallonne sous le contrat 01/1/4675 – http ://www.galilei.ulb.ac.be

Comme pour tout projet de recherche documentaire un mode de représentation des documents est nécessaire. C'est donc dans l'optique d'une description de profil pour l'application de groupement que la méthode présentée dans cet article a été développée.

L'objectif de celle-ci est la substitution de l'espace vectoriel des termes par un espace vectoriel des concepts plus compact.

2. État de l'art

L'augmentation continue de la masse d'informations numérique a rendu son traitement informatisé inévitable. Pour le rendre possible, différents modèles de représentation des textes (vectorielle, probabiliste, connexionniste, ...) adaptés au mode de traitement informatique ont été développés. À titre d'exemple, on peut citer le modèle vectoriel standard (Salton et McGill, 1983), les modèles LSI (Deerwester *et al.*, 1990), SOM (Self Organizing Map),...

Pour permettre l'utilisation d'un modèle, l'indexation des documents est nécessaire. Celle-ci introduit toute une série de problèmes. Pour tenter d'y répondre, de nombreuses méthodes ont été développées. La segmentation d'un texte en unité linguistiques peut s'effectuer à l'aide de règles (tokenisation) ou sur base d'un lexique (lexematisation). Toutefois certains séparateurs ambigus — comme les mots composés, les synonymes ou encore les acronymes — rendent la tâche ardue et nécessite un grand nombre de traitements supplémentaires. Ainsi les ambiguïtés lexicales et sémantiques peuvent être en grande partie résolues par des méthodes d'étiquetage morpho-syntaxique (Brill, 1992 ; Church, 1988 ; Cutting *et al.*, 1992 ; Roche et Schabes, 1995) et de désambiguïsation sémantique (Krovetz et Croft, 1992), la variabilité des mots par des techniques de désuffixation (Paternostre *et al.*, 2002 ; Porter, 1980), l'utilisation d'un dictionnaire ou l'utilisation de règles de dérivation fondées sur une analyse linguistique (Tzoukermann et Jacquemin, 1997).

Il existe aussi différentes techniques permettant de prendre en compte l'importance relative des unités linguistiques. L'utilisation d'anti-dictionnaires permet l'élimination des termes nécessaires à la construction de la phrase, mais généralement sans contenu informationnel. L'extraction de l'information la plus pertinente peut être réalisée à l'aide de filtres fréquentiels ou encore par l'utilisation des informations apportées par une analyse syntaxique.

Parmi l'ensemble de toutes ces méthodes des choix doivent être effectués en fonction des objectifs recherchés.

3. Représentation de l'information dans GALILEI

Les impératifs de ce projet sont ceux d'un système réel, avec toutes les limitations qui y sont associées : temps de calcul, place mémoire, autant de paramètres à prendre en compte et à minimiser. C'est dans ce cadre que la méthode a été développée à partir de la représentation déjà disponible.

Les documents analysés se décomposent en unités linguistiques plus ou moins porteuses de sens. Afin d'alléger la quantité de termes différents à stocker, des traitements préalables ont été appliqués.

Certains mots courants, nécessaires à la construction de la phrase, n'apportent aucune information sur la nature du sujet traité par le document. En absence de traitement morpho-syntaxique, ces mots peuvent être supprimés par l'utilisation d'un anti-dictionnaire (*stoplist*). Un deuxième traitement est appliqué afin de supprimer la trop grande variabilité des mots. En effet les variabilités flexionnelles (pluriel, conjugaison) et les variabilités dérivationnelles (passage d'une

catégorie morpho-syntaxique à une autre) introduisent un grand nombre de termes différents raccordés à une même racine et donc à un même sens. Les techniques de désuffixation permettent de supprimer pour une bonne part ces variations morphologiques (algorithme de Porter pour l'anglais (Porter, 1980), Carry pour le français (Paternostre *et al.*, 2002)). Le dictionnaire formé suite à cette phase d'analyse sera donc composé de radicaux.

4. Méthode

La méthode présentée dans cet article a pour objectif de regrouper plusieurs termes significatifs en entitées représentatives que l'on qualifie de « concepts ». Cette association pourra être envisagée par la seule connaissance du contenu d'un ensemble de documents et devra permettre une réduction de l'espace de représentation tout en gardant le caractère discriminant de celle-ci.

Cette opération sera réalisée sur base de relations de similitude entre les différents termes au niveau de leur représentation dans les documents et de leur contexte d'apparition, mais aussi de leur spécificité et de leur capacité de description. La réalisation d'un espace conceptuel selon la méthode présentée peut se diviser en plusieurs étapes.

Dans un premier temps, les termes du dictionnaire susceptibles d'être groupés sont sélectionnés afin d'alléger le traitement (diminution du temps de calcul et de la capacité de stockage nécessaire). Sur base de ceux-ci l'algorithme peut être appliqué et des concepts sont ainsi formés. Enfin la transition vers le nouvel espace de représentation doit être effectuée pour les différents types d'information tels que les documents. Les différentes étapes sont donc :

1. La sélection des termes utilisés par l'algorithme.
2. La création des concepts.
3. La représentation de l'information sur base du nouvel espace créé.

4.1. *Sélection des termes*

En complément de l'anti-dictionnaire et des algorithmes de désuffixation, un filtre fréquentiel est utilisé pour ne garder que les termes les plus discriminants avant de leur appliquer l'algorithme proprement dit. Ce filtre éliminera ainsi les termes présents dans un nombre de documents trop faible ou trop élevé (ces nombres resteront proportionnels au cas étudié : nombre de documents et de catégories. Dans le cas des tests réalisés, ces limites sont de l'ordre de [10,300]).

4.2. *Création des concepts*

La création des concepts est réalisée par regroupement des unités linguistiques. Ce regroupement est effectué sur base de la comparaison des différentes caractéristiques de chaque couple de termes. Parmi ceux-ci, les plus proches sont rassemblés pour former les groupes.

Dans un premier temps, les termes sélectionnés sont classés selon l'ordre décroissant de leur facteur IDF (Inverse Document Frequency).

$$IDF_i = log\frac{|D|}{|D_i|} \tag{1}$$

$|D|$ = Nombre total de documents.
$|D_i|$ = Nombre total de documents contenant le terme i.

Cet indice permet de déterminer la capacité de chaque terme à représenter un petit groupe de documents. En effectuant un classement selon l'ordre décroissant, les premiers traités seront

ceux catégorisant le mieux l'information. Cela permet d'éviter de grouper en priorité les termes plus généraux présents dans de nombreux documents.

Ils seront ensuite comparés à tous les éléments de la liste et, sur base de leurs similitudes et des groupes déjà créés, insérés dans un groupe.

Les différentes caractéristiques à la base de la comparaison sont les suivantes :

1. Confiance ;
2. Similarité de voisinage ;
3. Cooccurrence.

4.2.1. *Confiance*

La confiance (Maedche et Staab, 2000) d'un terme au sein d'un couple exprime le nombre de documents dans lesquels les deux éléments du couple sont présents par rapport au nombre de documents dans lesquels ce terme apparaît.

$$confiance(i_i^k \Rightarrow i_j^k) = \frac{|\{d_k | i_i^k \cup i_j^k \subseteq d_k\}|}{|\{d_i | i_i^k \subseteq d_i\}|} \tag{2}$$

avec i_i^k et i_j^k, les deux termes composant le couple k.

À chaque terme est associé un certain nombre de documents, ainsi la confiance par rapport à un terme permet de décrire la part de ses documents décrite par un couple l'incluant. Chaque couple sera représenté par sa confiance minimum ainsi que par sa confiance moyenne.

$$confiance_{min} = min\{confiance(i_i^k \Rightarrow i_j^k), confiance(i_j^k \Rightarrow i_i^k)\} \tag{3}$$

$$confiance_{moy} = \frac{1}{2} \cdot (confiance(i_i^k \Rightarrow i_j^k) + confiance(i_j^k \Rightarrow i_i^k)) \tag{4}$$

4.2.2. *Similarité de voisinage*

Cette mesure se base sur la comparaison des contextes d'apparition de deux termes. Pour cela, à chaque terme i est associé un vecteur $\vec{V_i}$ des termes voisins du premier à travers tous les documents décrits par celui-ci. Ce vecteur est obtenu par déplacement d'une fenêtre de taille k à travers l'ensemble des documents analysés. La mesure de la similarité peut alors être obtenue par le produit scalaire normé des vecteurs de voisins des deux termes.

$$sim_v(i_i, i_j) = \frac{\vec{V_i} \cdot \vec{V_j}}{\|\vec{V_i}\| \cdot \|\vec{V_j}\|} \tag{5}$$

Cette mesure correspond au calcul de la matrice d'association scalaire à la différence que celle-ci ne se compose plus de tous les termes du dictionnaire (sélection) et que le voisinage comparé n'a plus la taille du document, mais celui d'une fenêtre de dimension k. Signalons encore que lors de l'analyse du document au moyen de la fenêtre de taille k tous les termes sont encore présents, lors de la construction du vecteur de voisinage ce n'est plus le cas. Ainsi chaque occurrence du terme n'apportera pas (k-1) voisins, mais bien

$$\{v_i\} \setminus \{\{v_i \in stoplist\} \cup \{v_i | |D_i| \notin [min, max]\}\} \tag{6}$$

où D_i est l'ensemble de tous les documents dans lesquels apparaît le terme v_i.

4.2.3. *Cooccurrence*

La cooccurrence de deux termes correspond au nombre d'apparitions simultanées de ceux-ci dans l'ensemble des documents, ou plus simplement le nombre de documents contenant les deux termes. Elle s'obtient en effectuant le produit scalaire des vecteurs documents de chacun des deux termes.

4.2.4. *Comparaison*

Il est évident que dans une certaine mesure ces différents facteurs sont liés (ex : deux termes ayant une cooccurrence nulle auront de la même manière des indices de confiance nuls). Toutefois chacune de ces mesures a sa propre spécificité. Il est en effet inutile de grouper deux termes n'apparaissant jamais simultanément dans un même document (cooccurrence) tout en n'étant pas exclu que bien qu'apparaissant peu dans les mêmes documents, ils soient relatifs à un même concept (similarité de voisinage). De la même façon, la confiance permet de se rendre compte du manque de précision de la mesure de cooccurrence car certains couples de cooccurrence élevée peuvent obtenir une confiance de l'ordre de 100 % pour un terme et ridiculement basse pour l'autre, indiquant par là la réunion d'un terme spécifique avec un terme très général. L'utilisation simultanée de ces différentes informations doit pouvoir permettre de remédier aux différentes difficultés rencontrées.

Le couple retenu sera celui répondant à la condition suivante :

$$\max_j \{confiance_{min}(i_i, i_j) + confiance_{moy}(i_i, i_j) + sim_v(i_i, i_j)$$

$$+(confiance_{min}(i_i, i_j) \cdot confiance_{moy}(i_i, i_j) \cdot sim_v(i_i, i_j))\} \quad (7)$$

avec des valeurs minimum pour la $confiance_{min}$ et la $cooccurrence$.

Ainsi, après formation et ordonnancement de la liste des termes à grouper, celle-ci est traitée itérativement terme après terme dans l'ordre de la liste. À chaque itération, le terme est comparé à tous les autres et est soit inséré dans un groupe déjà existant, soit un nouveau groupe est créé, suivant le résultat de la comparaison. Si aucun couple ne satisfait les conditions, le terme est supprimé de la liste.

4.3. *Représentation de l'information*

Le nouvel espace conceptuel étant formé, la translation d'un espace vectoriel à un autre peut être réalisée. Les documents ne seront plus représentés que par les concepts qu'ils contiennent. Un concept sera considéré comme apparaissant dans un document lorsqu'un minimum de deux des termes qu'il représente sont contenus dans le document. La fréquence associée à ce concept sera la somme de l'ensemble des termes associés au concept apparaissant dans le document.

5. Tests

Différents testes ont été effectué sur différentes bases de données pré-catégorisées : News1500 (anglais, 1480 documents répartis en 13 catégories), News20000 (anglais, 19778 documents répartis en 24 catégories), LeSoir (français, 13234 documents répartis en 9 catégories). Des bases de données pré-catégorisées sont utilisées pour permettre une division en profils et la simulation des jugements.

5.1. *Mesures*

L'objectif est de représenter l'information de manière plus compacte afin de diminuer l'espace de stockage et le temps de calcul. Les premières mesures à vérifier sont donc la variation du nombre de dimensions de l'espace de représentation et la qualité de la représentation.

Tant dans l'espace vectoriel des mots que dans l'espace des « concepts », les documents sont représentés par un vecteur $\vec{d_j} = (d_{1,j}, d_{2,j}, ..., d_{t,j})$ dont les composantes sont calculées de la manière suivante (Salton et McGill, 1983) :

$$d_{i,j} = \frac{o_{i,j}}{\max_l o_{l,j}} \cdot log \frac{|D|}{|D_i|} \tag{8}$$

$o_{i,j}$ = occurrence du terme i dans le document j.

La similarité entre deux documents sera calculée en effectuant le produits scalaire normé de deux vecteurs (similarité cosinus)

$$sim(\vec{d_i}, \vec{d_j}) = \frac{\vec{d_i} \cdot \vec{d_j}}{\|\vec{d_i}\| \|\vec{d_j}\|} \tag{9}$$

Similarité interne moyenne : La similarité interne est la similarité du vecteur d'un document par rapport au vecteur moyen de la catégorie à laquelle il appartient.

$$\frac{1}{|D|} \cdot \sum_{d_i \in D} sim(\vec{d_i}, \vec{d}^{k_i}) \tag{10}$$

$$\vec{d}^{k_i} = \frac{1}{|D_{k_i}|} \cdot \sum_{d_i \in D_{k_i}} \vec{d_i} \tag{11}$$

$\vec{d}^{k_i}$= vecteur moyen de la catégorie k_i à laquelle appartient le document i.

Similarité externe moyenne : La similarité externe est la similarité du vecteur d'un document par rapport au vecteur moyen de la catégorie à laquelle il n'appartient pas et dont il est le plus proche.

$$\frac{1}{|D|} \cdot \sum_{d_i \in D} \max_{j \neq i} \{sim(\vec{d_i}, \vec{d}^{k_j})\} \tag{12}$$

Pourcentage de documents bien catégorisés : Pourcentage de documents dont la similarité interne est plus élevée que la similarité externe. Ce sont donc les documents pour lesquels le vecteur est le plus proche du vecteur moyen de la catégorie à laquelle ils appartiennent.

Ces différentes mesures permettrons de vérifier que les différents sujets sont bien identifiés et que les profils pourront ainsi être groupés.

5.2. *Simulation*

L'objectif étant d'intégrer la méthode dans GALILEI, des simulations ont été effectuées sur les différentes bases de données. Pour simuler un système réel où les profils sont groupés en communautés virtuelles, la procédure utilisée est décrite ci-après.

On part d'une base de données catégorisée. Deux profils attachés chacun à une des catégories sont créés. Un ensemble de jugements sur les documents est simulé : un document appartenant à la même catégorie que le profil avec lequel il est jugé, est considéré comme pertinent, dans l'autre cas, comme hors sujet. Ils sont alors groupés en communautés virtuelles et ce résultat est comparé à la solution idéale.

Ensuite, se succèdent au fil du temps des phases de jugement des documents partagés au sein des communautés virtuelles et des phases de création d'un nouveau profil associé à un ensemble de jugements de documents.

Le test porte sur 500 unités de temps. Chaque unité de temps correspondant à une phase de jugement ou une phase de création de nouveau profil. À la fin de chaque unité de temps un nouveau groupement est calculé et la solution réévaluée. L'évaluation de la solution se fait à travers l'*Adjusted RAND Index* (Hubert et Arabie, 1985) qui représente la qualité de la réponse par rapport au groupement idéal. Cette valeur est comprise entre -1 et 1. Lorsque celle-ci est nulle, le groupement calculé est aléatoire.

6. Résultats

6.1. *Exemples de mots groupés (LeSoir) :*

douma - kremlin
medical - infirmier - hopital
gerhard - schröder
asterix - goscinny - uderzo
afghanistan - taliban - kaboul
échappement - rejet - consommation - pollution
...

On peut voir à partir de ces quelques exemples de groupements réalisés pour la base de données LeSoir, la cohérence des groupes créés.

6.2. *Mesure de similarité*

Data Base	langue	Représentation	similarité interne	similarité externe	pourcentage	dimension
News1500	en	mots	0.38302946	0.24871015	86.0135	38914
		concepts	0.37456777	0.21177723	88.0405	2304
News20000	en	mots	0.28735075	0.21038529	71.9486	109965
		concepts	0.19984695	0.08734833	74.6890	5590
LeSoir	fr	mots	0.60287311	0.60384188	60.7520	100471
		concepts	0.18022144	0.08921644	85.3786	5821

Tableau 1. Résultats

Les résultats permettent de constater une nette diminution de la taille de l'espace de représentation (rapport supérieur à 15) tout en apportant une différentiation entre catégories légèrement meilleure (+2 ou 3% pour l'anglais et +25% pour le francais).

6.3. *Simulation*

Les résultats des simulations sur les différentes bases de données montrent bien qu'à partir d'une certaine masse d'informations le système fonctionne de manière équivalente dans les deux

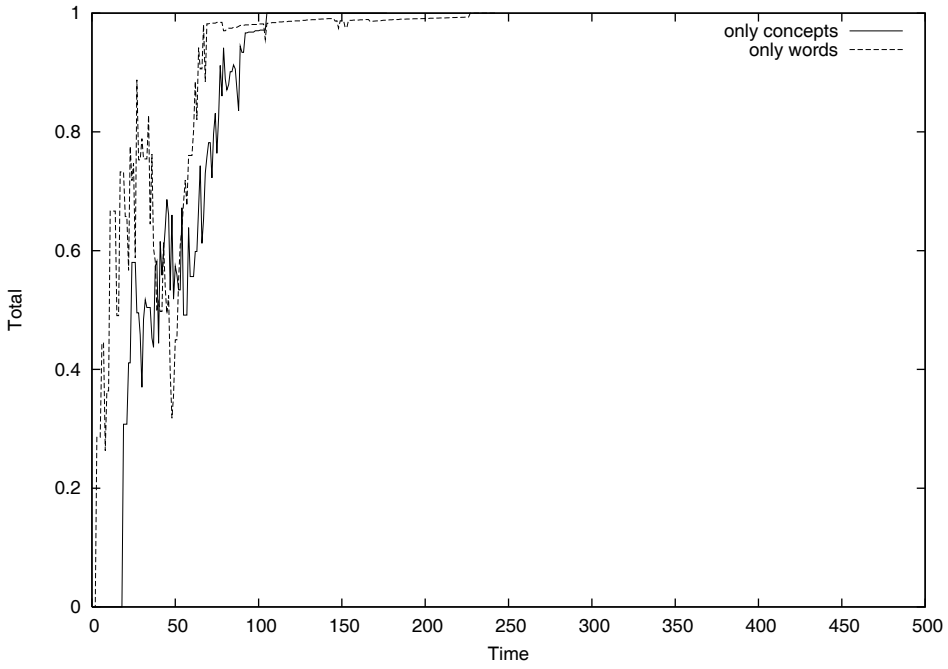


Figure 1. Simulation d'un système réel sur News1500

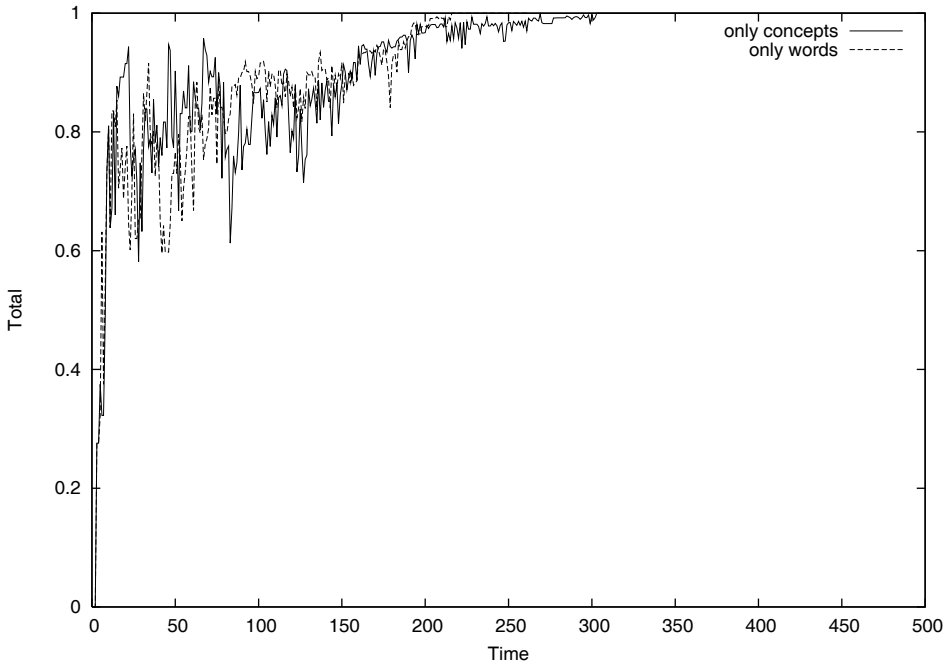


Figure 2. Simulation d'un système réel sur News20000

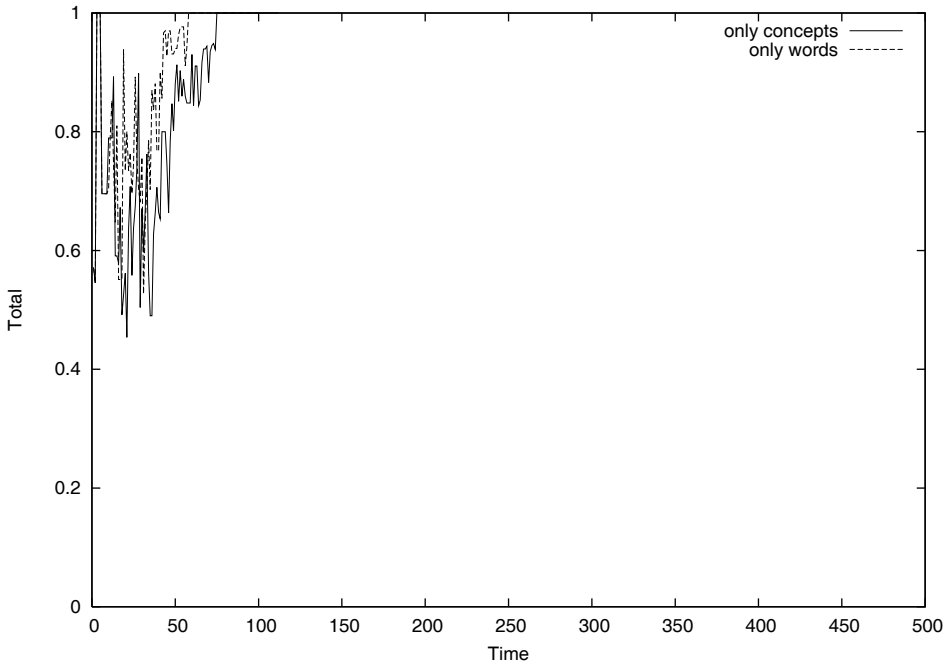


Figure 3. Simulation d'un système réel sur LeSoir

modes de représentation à la différence près que dans l'espace conceptuel le temps nécessaire au calcul est nettement moindre. (Figure 1, Figure 2, Figure 3)

7. Conclusion

L'objectif de la méthode proposée était la génération automatique d'un système de représentation valide et compact, dans l'optique d'un traitement de l'information plus rapide et moins gourmand en ressources. Les résultats obtenus montrent une diminution conséquente des dimensions sans perte importante d'information. La réduction des dimensions de l'espace de représentation sur base de l'analyse des données contextuelles a donc bien été réalisée.

L'utilisation d'une telle méthode permet donc un gain de temps relativement important au niveau des différents traitements appliqués à l'information ainsi qu'une diminution de la capacité de stockage nécessaire, tout en n'atténuant pas la qualité de la représentation. De plus, un de ces principaux avantages est son entière automatisation.

Références

Brill E. (1992). A Simple Rule-Based Part-of-Speech Tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing* : 152-155.

Church K.W. (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the 2nd ACL Conference on Applied Natural Language Processing* : 136-143.

Cutting D., Kupiec J., Pedersen J. et Sibun P. (1992). A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing* : 133-140.

Deerwester S., Dumais S., Landauer T., Furnas G. et Harshman R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, vol. (41/6) : 391-407.

Francq P. (2003). *Stuctured and Collaborative Search : An integrated approach to share documents among users*. PhD Thesis. Université Libre de Bruxelles.

Hubert L. et Arabie P. (1985). Comparing partitions. *Journal of Classification* : 193-218.

Krovetz R. et Croft W.B. (1992). Lexical ambiguity and information retieval. *Information Systems*, vol. (10/2) : 115-141.

Maedche A. et Staab S. (2000). Discovering conceptual relations from text. In *Proceedings of the 13th European Conference on Artificial Intelligence* : 321-325.

Paternostre M., Francq P., Saerens M., Lamoral J. et Wartel D. (2002). Carry, un algorithme de désuffixation pour le français. Version électronique disponible sur [http ://www.galilei.ulb.ac.be].

Porter M.F. (1980). An Algorithm for Suffix Stripping. *Program*, vol. (14/3) : 130-137.

Roche E. et Schabes Y. (1995). Deterministic Part-of-Speech Tagging with Finite-State Transducers. *Computational Linguistics*, vol. (21/2) : 227-253.

Salton G. et McGill M. (1983). *Modern Information Retrieval*. McGraw-Hill Book Co.

Tzoukermann E. et Jacquemin C. (1997). Analyse Automatique de la Morphologie Dérivationnelle et Filtrage de Mots Possibles. In *Actes du Colloque Mots possibles et mots existants* : 251-260.

Modélisation de l'Hyperonymie via la combinaison de réseaux sémantiques et de vecteurs conceptuels

Mathieu Lafourcade, Violaine Prince

LIRMM – Montpellier – France
{lafourcade, prince}@lirmm.fr

Abstract

Lexical semantics are a key issue in Natural Language Processing since they represent the point of convergence with conceptual Knowledge Representation and ontologies. Among the well established trends in lexical semantics representations, two trends seem to be conflictual : the WordNet approach, born from semantic networks, and KR-oriented, and the Vector approach, originated from the Saltonian representation in Information Retrieval, which has found a set of applications in NLP. When ontological representation is needed, hypernoymy, the closest approximation to the *is-a* relation, is at stake. We show how to account for hyperonymy within the vector-based frame for semantics, relying on a cooperation between semantic networks and conceptual vectors.

Résumé

La sémantique lexicale est un élément clé du Traitement Automatique du Langage Naturel en ce qu'elle représente le point de convergence entre les représentations des connnaisances et les ontologies. Parmi les tendances durables en représentation de sémantique lexicale, deux d'entres-elles semblent être conflictuelles ; l'approche WordNet issue des réseaux sémantiques et teintée de KR, et l'approche vectorielle issue des représentations saltoniennes et de la recherche d'information qui a su trouver un large spectre d'applications en TALN. Quand les représentations ontologiques sont nécessaires, l'hyperonymie, l'approximation la plus proche de la relation *is-a* , est en jeu. Nous montrons comment gérer l'hyperonymie dans le cadre d'une approche vectorielle pour la sémantique, en combinant les réseaux sémantiques et les vecteurs conceptuels.

Mots-clés : traitement automatique du langage naturel, apprentissage lexical, vecteurs conceptuels, hyperonymie, réseaux sémantiques.

Keywords: natural language processing, lexical learning conceptual vectors, hyperonymy, semantic networks.

1. Introduction

La sémantique lexicale est un aspect important du Traitement Automatique du Langage Naturel (TALN) en ce qu'elle constitue le point de convergence entre les représentations des connaissances (KR) conceptuelles et les ontologies. Elle s'étend largement au domaine du traitement des ressources lexicales, et de nombreux travaux à la fois en TALN et en IA ont été dédiés aux fonctions sémantiques lexicales comme un moyen d'approcher la représentation des sens des mots et leur discrimination. Parmi les tendances classiques en representation lexicale et sémantique, deux approches semblent conflictuelles : celle de Wordnet (Miller et Fellbaum, 1991 ; Fellbaum, 1998) issue des réseaux sémantiques et teintée de KR, et les approches vectorielles issues des représentations saltoniennes en recherche d'information (Salton, 1968), qui ont su trouver de nombreuses applications en TALN. La première est globalement basée sur des lo-

giques et la seconde sur des algèbres d'espaces vectoriels. La première est très efficace pour les relations du type *is-a* (considéré comme une relation conceptuelle souvent incluse dans l'hyperonymie), mais demeure pratiquement muette à propos d'autres fonctions lexicales intéressantes comme l'antonymie ou les associations thématiques. La synonymie a été largement traitée (Resnik, 1995 ; Sparck Jones, 1986), mais la discrimination entre synonymie et hyperonymie a souvent amené à étudier des notions plus flexibles comme la similarité sémantique (Resnik, 1999). L'approche vectorielle se situe complètement à l'opposé. Permettant naturellement les associations thématiques, elle s'offre à l'implémentation de nombreuses fonctions de synonymies à grain fin (Lafourcade et Prince, 2001) mais aussi de fonctions d'antonymie (Schwab *et al.*, 2002) tout en restant incapable de différencier ou de valider l'existence de relations d'hyperonymie. Dans cet article, nous montrons (1) comment modéliser l'hyperonymie dans le cadre vectoriel pour la sémantique, en se basant sur une coopération entre les réseaux sémantiques et les vecteurs conceptuels, (2) comment elle peut être appliquée à d'autres nouvelles fonctions comme la substitution de termes et l'approximation sémantique qui toutes les deux font partie du champ de la similarité sémantique. Nous utilisons un réseau sémantique afin d'améliorer l'apprentissage de vecteurs, et de façon symétrique, nous construisons un sous-réseau spécifique de relations d'hyperonymie entre vecteurs sémantiques.

2. Hyperonymie, relations *is-a* et vecteurs conceptuels

L'**hyperonymie** est une fonction lexicale qui, à partir d'un terme t, associe un ou plusieurs autres termes plus généraux, comme ceux utilisé pour définir t en *genre* (définition aristotélicienne). Les définitions de dictionnaire sont en général en *genre* et en *différences*. L'hyperonymie, dans la plupart des articles en KR, est assimilée à la relation *is-a* (voir Brachman et Schmolze, 1985). Rappelons ici que la relation *is-a* est telle que si X est une classe d'objets, et X' une sous-classe de X, alors is-a(X', X) est vrai. Le terme X est l'argument général tandis que X' est l'argument spécifique. Le problème que nous rencontrons est que l'hyperonymie linguistique n'est pas une relation *is-a* "pure" (héritage simple). Nous trouvons, par exemple, la définition suivante du terme ‹*cheval*› : "un animal herbivore, quadrupède, etc." Un bon hyperonyme pour cette définition de ‹*cheval*› est un ‹*animal herbivore*› ou encore ‹*quadrupède herbivore*›. Le terme ‹*animal*› est un autre hyperonyme, tandis que la relation ‹*mammifère herbivore*› *is-a* ‹*mammifère*› est vrai. Cependant, thématiquement, un ‹*cheval*› est très proche d'un ‹*herbivore*›, alors que ‹*herbivores*› est un ensemble d'individus qui peuvent appartenir à plusieurs sous-hiérarchies d'une taxonomie (certains oiseaux, insectes et reptiles sont herbivores, mais aussi métaphoriquement plein d'autres choses). Donc, même si linguistiquement on peut toujours écrire qu'*un cheval est un herbivore*, il demeure que la relation ‹*cheval*› *is-a* ‹*herbivore*› pose problème. Plusieurs solutions sont alors envisageables, dont en premier lieu considérer que l'héritage multiple de *is-a* est possible. Si tel n'est pas le cas, alors si on choisit ‹*cheval*› *is-a* ‹*mammifère*›, il semble nécessaire de représenter les autres relations (‹*cheval*› *is-a* ‹*herbivore*›, ‹*cheval*› *is-a* ‹*quadrupède*›, etc.) comme des propriétés (ou des rôles ou des vues selon la terminologie utilisée dans la modélisation concernée). En ce qui nous concerne, l'héritage multiple de classes, motivées à la fois par leur utilité (factorisation) et leur interprétation naturelle, nous semble la plus adaptée.

Wordnet et l'hyperonymie. Wordnet a pour relation fondamentale la synonymie. Cependant, en ce qui concerne l'hyperonymie, Wordnet s'apparente à une taxonomie de termes, et en tant que telle, elle ne capture que les relations *is-a* . Un hyperonyme est un super-ordonné linguistique, généralement utilisé dans des définitions, qui capture également certaines propriétés qui ne peuvent pas jouer le rôle de classes par elles-mêmes. Les termes polysémiques ont, en général, de nombreuses définitions et donc souvent beaucoup d'hyperonymes : un ‹*cheval*› est

également un ‹*véhicule*›, ce qui explique pouquoi WordNet est organisé selon un réseau et non pas un arbre. La seule contrainte langagière est que l'hyperonyme doit être plus général (et donc ‹*herbivore*› peut être un hyperonyme pour ‹*cheval*›) alors que dans un réseau sémantique, chaque étape de la chaînes de classes et sous-classes doit vérifier la relation d'ordre.

Hyperonymie et définition de termes. Comme évoqué précédemment, les hyperonymes peuvent être extraits, quand ils sont inconnus, de définitions dictionnairiques. Seuls les concepts généraux, qui ont tendance à jouer le rôle d'hyperonymes et de superclasses (*is-a*) de nombreux termes, ne sont pas définis via une approche aristotélicienne, mais sont appréhendés à travers leurs hyponymes. C'est pourquoi, dans notre modèle de vecteurs conceptuels (présenté dans la section suivante), nous considérons l'existence d'un *horizon d'hyperonymie* au-dela duquel les définitions doivent être inversées : les hyperonymes sont plus difficiles à trouver et sont moins explicatifs que les hyponymes. Le terme ‹*action*› est pratiquement au sommet de la taxonomie de WordNet et les définitions de dictionnaires tendent à l'expliquer avec des termes plus spécifiques.

Vecteurs conceptuels. Les vecteurs ont été utilisés depuis longtemps en Recherche d'Informations (Salton et MacGill, 1983) mais aussi pour la représentation du sens dans LSI issu des études d'analyse sémantique latente (LSA) en psycholinguistique (Deerwester *et al.*, 1970). En TALN, Chauché (1990) a proposé un formalisme pour la projection de la notion linguistique de champ sémantique dans un espace vectoriel, sur lequel notre modèle se fonde. À partir d'un ensemble de notions élémentaires, *les concepts*, il est possible de construire des vecteurs (dits conceptuels) et de les associer à des items lexicaux[1]. L'hypothèse qui considère qu'un ensemble de concepts peut être un générateur pour le langage a depuis longtemps été exposée dans Rodget (1852) (hypothèse du thésaurus) et a été appliqué par les chercheurs en TALN (par exemple Yarowsky, 1992). Les termes polysémiques combinent les différents vecteurs correspondant aux différents sens. Cette approche vectorielle est basée sur des propriétés mathématiques connues, et il est donc possible d'effectuer des manipulations formelles auxquelles sont attachées des interprétations linguistiques (ou psycho-linguistiques) raisonnables. Les concepts sont définis par un thésaurus (dans notre prototype appliqué au français, nous avons choisi Larousse (2001) où 873 concepts sont identifiés — on peut comparer avec le millier de concepts définis dans Rodget, [1852]). Afin d'être cohérent avec l'hypothèse du thésaurus, on considère que cet ensemble constitue un espace générateur pour les mots et leurs sens. Cet espace n'est pas libre (pas de base propre) et n'importe quel terme peut y projeter ses sens selon le principe suivant.

Soit $\mathcal{C}$ un ensemble fini de n concepts, un vecteur conceptuel V est une combinaison linéaire des élements c_i de $\mathcal{C}$. Pour un sens A, son vecteur $V(A)$ est la description en extension des activations de tous les concepts de $\mathcal{C}$. Par exemple, les différents sens de ‹*porte*› peuvent être projetés sur les concepts suivants (avec leur valeur d'activation entre crochets) : V(‹*porte*›) = (*OUVERTURE*[0.7], *BARRIÈRE*[0.41], *LIMITE*[0.35], *PROXIMITÉ*[0.31], *EXTERIEUR*[0.30], ...

En pratique, plus $\mathcal{C}$ est grand, plus fine est la description du sens. En retour, la manipulation informatique est moins aisée. Comme les vecteurs sont globalement denses (très peu de composantes valent 0), l'énumération des concepts activés est longue et passablement difficile à évaluer. On préfère, en général, sélectionner les termes thématiquement les plus proches, c'est-à-dire son *voisinage*. Par exemple, les termes les plus proches de ‹*porte*› ordonnés par distance thématique croissante sont : $\mathcal{V}$(‹*porte*›) = ‹*portillon*›, ‹*portail*›, ‹*portière*›, ‹*issue*›, ‹*ouverture*›, ...

[1] Un item lexical est un terme simple ou composé, ou encore une locution. Par exemple ‹*voiture*›, ‹*pomme de terre*› et ‹*tirer le diable par la queue*› sont des items lexicaux.

Afin de représenter et manipuler certains aspects de sémantique lexicale dans ce cadre vectoriel, nous utilisons des opérations élémentaires sur les vecteurs. Une mesure intéressante est la distance angulaire qui se base sur une mesure classique de similarité. Une approche incluant l'hyperonymie s'avère particulièrement pertinente pour améliorer la construction des vecteurs, car la plupart de ceux-ci est construite via l'analyse des définitions hyperonymiques (disponibles sur le Web ou dans des dictionnaires en ligne). En pratique, nous avons extrait environ 250000 liens pondérés d'hypéronymie pour environ 40000 substantifs (soit 120000 acceptions). En fait, l'ensemble des fonctions lexicales semble être particulièrement utile pour une telle tâche. De façon symétrique, les relations entre les vecteurs sont cruciales pour une telle approche guidée par les données : essayer d'extraire les relations sémantiques de corpus (Yarowsky, 1992) et ainsi construire une ontologie d'un domaine, ou essayer d'organiser l'information des corpus en se basant sur des hiérarchies *is-a* (Lee *et al.*, 1993 ; Resnick, 1999).

3. Le calcul de l'hyperonymie

Notre approche étant à la fois guidée par les données et basée sur des hiérarchies, nous définissons deux mesures qui sont utilisée conjointement. Le modèle de co-occurrence permette de formaliser la sustitution de termes et l'approximation sémantique (avec un aspect taxonomique). Le modèle d'inclusion (de propriétés) directement basé sur les vecteurs conceptuels : une sous-classe contient les propriétés de sa superclasse. La combinaison de ces deux modèles permet de contruire des réseaux sémantiques partiels qui constitue une *mémoire incrémentale* qui instancie pour des termes les mesures de substitution et d'approximation sémantique. Conjointement à leur construction, ces réseaux sémantiques sont aussi utilisés comme support pour la propagation et l'affinage de vecteurs conceptuels.

3.1. *Modèle de co-occurrence*

On définit deux mesures entre un terme t et un **candidat hyperonyme** h :

$$M_S(t,h) = \frac{|H \cap T|}{|T|} \quad \text{et} \quad M_T(t,h) = \frac{|H \cap T|}{|H|} \tag{1}$$

T (resp. H) représente l'ensemble de documents dans un corpus donné, qui contiennent le terme t (resp. h). $H \cap T$ représente l'ensemble de documents contenant les deux termes h et t. M_S se rapproche de la notion de *rappel* et M_T de celle de *précision* en Recherche d'Information.

Substitution de termes et approximation sémantique : vers une hiérarchie *is-a* locale.

Si on ajoute l'hypothèse que *h est un hyperonyme possible*, alors M_S évalue dans quelle mesure t peut être subsitué par h et consitue ce que nous appelerons une mesure de substitution. De façon similaire, M_T est une mesure d'évaluation de taxonomie, et correspond à une *approximation sémantique* (de la même façon que l'on peut approximer ‹*cheval*› par ‹*mammifère*›).

Nous avons conduit une expérimentation en utilisant Google *(www.google.com)* et le nombre de documents trouvés pour chaque requête. Statistiquement, cette approximation semble largement compensée par la grande taille de corpus que représente le Web. Par exemple, nous avons les résultats suivants pour le terme ‹*avion*› :

aéronef /$M_T = 0.26, M_S = 0.025$ | appareil volant /$M_T = 0.53, M_S = 0.0007$
appareil /$M_T = 0.12, M_S = 0.17$ | aéronef plus lourd que l'air /$M_T = 0.52, M_S = 0.00004$

Le meilleur terme de substitution est ‹*appareil*› (plus élevée M_S), mais c'est également la pire des approximations sémantiques. Le terme ‹*appareil volant*› est le plus précis (M_T la plus forte), mais ne peut pas raisonnablement être utilisé à la place d'‹*avion*›.

Dans le cas de ‹*cheval*›, nous créons un nouveau sens (‹*cheval/moyen de transport*› et ‹*cheval/ viande*›) et les lions à leur hyperonymes respectifs. Le problème est qu'en commençant à partir de définitions vectorisées, il n'y a pas moyen d'appréhender ces nouveaux sens car ils ne sont pas (encore) identifiés. Pour surmonter cette difficulté, nous lions chacun de ces nouveaux sens au sens le plus proche (selon la distance thématique) déjà existant. Dans l'exemple précédent, nous avons :

- ‹*cheval/ moyen de transport*› est plus proche de ‹*cheval/mammifère*› que de ‹*cheval/unité de puissance*›. Cette relation peut être vérifiée sur leurs vecteurs respectifs, et par la recherche de patrons lexicaux caractéristiques sur des définitions encyclopédiques.
- ‹*cheval/viande*› est plus proche de ‹*cheval/mammifère*› que de ‹*cheval/unité de puissance*›.

Ces deux mesures sont particulièrement utiles lors d'un processus d'analyse sémantique. En effet, la constitution d'un réseau lexical sur la base des deux mesures M_S et M_T permet de reconnaître les hyperonymes lointains de substitution (faible M_T et fort M_S). Par exemple, on peut lors de l'analyse s'appercevoir que la cohérence thématique du texte est bien plus forte lorsque l'on (re)substitue ‹*avion*› à ‹*appareil*›. Les candidats à substituer sont déterminés par la structure du réseau et la distance angulaire entre le candidat et le contexte global. Il s'agit d'un processus itératif globalement convergent. Ainsi, en vue de l'analyse, on procède à l'inverse de l'auteur, qui avait, directement ou non, remplacé les termes précis par des hyperonymes plus vagues à des fins stylistiques (par exemple, suppression des répétitions).

3.2. *Modèle d'inclusion*

On note $V(A)$ le vecteur conceptuel associé au terme A. Si A est un hyperonyme de B, alors les propriétés de A sont incluses dans celles de B. Cela peut-être mesuré via l'intersection de vecteurs et la distance angulaire :

$$H(A,B) \quad \Rightarrow \quad D_A(V(A), \gamma(V(A)\, V(B))) \quad \leq \quad D_A(V(B), \gamma(V(A), V(B))) \qquad (2)$$

Par exemple, nous avons les mesures suivantes entre ‹*cheval/mammifère*› et ‹*mammifère*› :

$$D_A(V(\text{cheval}), \gamma(V(\text{cheval})\, V(\text{mammifère}))) = 0.41$$
$$D_A(V(\text{mammifère}), \gamma(V(\text{cheval})\, V(\text{mammifère}))) = 0.25$$

De ces résultats, nous déduisons que les propriétés de ‹*mammifère*› sont incluses dans celles de ‹*cheval*›. De plus, si par ailleurs on sait que ‹*cheval*› et ‹*mammifère*› sont en relation d'hyperonymie, on en déduit que ‹*mammifère*› est l'hyperonyme. Bien sûr, pour connaître l'existence de cette relation entre deux termes, l'extraction de schémas caractéristiques (par exemple *X est un type de Y*) combiné au modèle de co-occurrence fournissent des indices forts.

3.3. *Limites des modèles*

Le modèle fonctionne correctement pour des vecteurs qui ont été calculés à partir de définitions hyperonymiques. L'évaluation se fait manuellement, ponctuellement pour les mesures sur certains termes, et globalement sur les réseaux sémantiques construits. Cependant, pour des termes très généraux, où les définitions ont tendance à être hyponymiques (une collection d'exemples), l'inclusion de vecteur est renversée. Plus précisement, la position limite où a lieu le renversement est appelée l'*horizon*. L'horizon est globalement constituée des concepts feuilles de la taxonomie sur laquelle est construit l'espace vectoriel des sens. Quand une définition conduit à un nouveau vecteur (ou à sa révision), les vecteurs des termes présents dans la définition sont mélangés. En conséquence, le vecteur résultat est relativement plat en comparaison des

principaux concepts impliqués. Nous avons une mesure formelle de *profil de vecteur* qui est le *coefficient de variation* CV :

$$\mathrm{CV}(X) = \frac{s(X)}{\mu(X)} \qquad \text{avec} \qquad s^2(X) = \frac{\sum_i (x_i - \mu(X))^2}{n} \tag{3}$$

Le terme CV est l'écart-type des composants x_i des vecteurs divisé par leur moyenne μ. C'est une valeur sans unité. Par définition, CV n'est défini que pour les vecteurs non nuls. Si $\mathrm{CV}(A) = 0$ alors A fait un angle de $\pi/4$ avec chacun des axes, et à sa valeur maximum (environ 29 quand $n = 873$), nous avons un vecteur booléen (une seule composante à 1 et toutes les autres à 0).

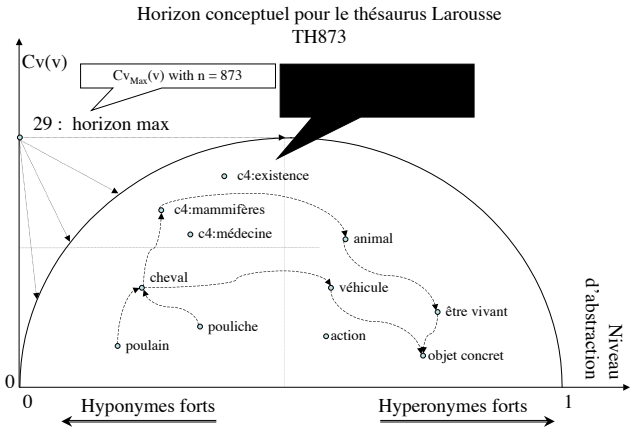


Figure 1. Représentation graphique en 2D de l'horizon conceptuel. L'horizon se situe au plus haut niveau du coefficient de variation (des vecteurs de la base) qui est le plus bas niveau de la hiérarchie du thesaurus. Sur le côté gauche, nous avons une généralisation des concepts, où de façon similaire, le mélange des vecteurs tend à faire baisser le coefficient de variation alors que le niveau d'abstraction augmente. Les termes préfixés par c4 : *correspondent aux concepts de niveau 4 définis dans le Larousse (2001).*

Au-dela de l'horizon, nous avons :

$$H(A,B) \quad \Rightarrow \quad D_A(V(A), \gamma(V(A)\, V(B))) \geq \quad D_A(V(B), \gamma(V(A), V(B))) \tag{4}$$

Comment évaluer de quel côté de l'horizon un vecteur donné se situe ? En lui-même, le coefficient de variation évalue uniquement la forme générale du vecteur et son taux de *conceptualité* relativement au jeu de concepts initial. Nous avons deux moyens de résoudre ce problème :

1. De se concentrer sur une approche lexicale combinant des fonctions et informations lexicales aux vecteurs. Le modèle de co-occurence constitue une réponse possible mais on peut également penser aux graphes conceptuels *à la Sowa*. C'est une approche multi-représentation dont la généralisation abusive risque toutefois d'être difficile à gérer.
2. D'inclure comme dimensions supplémentaires dans l'espace vectoriel, chaque concept de la hiérarchie et non pas seulement les concepts terminaux (les feuilles). Cette solution est clairement partielle, car elle ne peut pas être une réponse à celui de la polysémie si on se place au niveau du terme et non celui de l'acception.

4. Discussion et conclusion

Les expériences que nous avons conduites sur une collection de substantifs (et termes composés), ont mis en lumière le problème posé par l'horizon conceptuel. Cet horizon se situe au niveau le plus bas de la hiérarchie de concepts (nous avons utilisé le *Larousse* (2001) pour le

français). De par la nature de la composition de vecteurs, le modèle d'inclusion doit être inversé quand les termes se situent au-dela de cet horizon. La détection du passage de part et d'autre de l'horizon est réalisé par des modèles lexicaux. De façon plus précise, elle peut se baser sur le modèle de co-occurrences mais aussi sur l'identification des hyperonymes. La présentation détaillée de la découverte des hyponymes n'est pas l'objet de cet article (et les méthodes sont très classiques, voir Thelen et Riloff (2002) pour une approche récente d'extraction de catégories taxonomiques), cependant il est intéressant de noter que les termes les plus abstraits (qui correspondent à des classes taxonomiques de grande taille) contiennent un grand nombre d'hyperonymes. Selon notre modèle, les fonctions d'hyperonymie et d'hyponymie ne sont pas strictement symétriques (à la fois dans leur usage et leur comportement dans les corpus) et peuvent être utilisées conjointement pour renforcer la construction du réseau sémantique.

Nous avons essayé de montrer comment il est possible de modéliser l'hyperonymie dans le cadre d'une sémantique lexicale vectorielle, en se basant sur une coopération entre les réseaux sémantiques et le modèle des vecteurs conceptuels. Après avoir évalué l'importance, pour la sélection lexicale, de fonctions lexicales comme la synonymie et l'antonymie ainsi que la construction et l'utilisation des vecteurs conceptuels, nous nous sommes concentrés sur l'hyperonymie qui semble plus difficile à aborder sous un angle purement numérique. Notre approche étant à la fois guidée par les données et par la hiérarchie, nous avons d'abord chercher à définir l'impact de l'hyperonymie par des mesures dans des corpus. On peut ainsi définir formellement la substitution de termes et l'approximation sémantique (incluant l'aspect taxonomique). À partir du modèle théorique, qui combine réseaux sémantiques et vecteurs conceptuels, il a été possible de l'implémenter et de montrer comment l'inclusion a été traitée et quels résultats nous avons obtenus. Bien que satisfaisants, ces résultats tendent à refléter la nature polymorphe de l'hyperonymie : étant plus complexe qu'une simple relation *is-a* , l'hyperonymie doit être contrainte au sein de la tâche à entreprendre. Si l'enjeu est la correction ou l'explication de texte, alors la substitution de termes est une bonne utilisation des propriétés d'hyperonymie. Pour la construction d'une taxonomie, l'approximation sémantique est un meilleur candidat.

Références

Brachman R.J. et Schmolze J.G. (1985). An overview of the KL-ONE knowledge representation system. *Cognitive Science*, vol. (9/2) :171-216.

Chauché J. (1990). Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TA Information*, vol. (31/1) : 17-24

Deerwester S., Dumais S., Landauer T., Furnas G. et Harshman R. (1990). Indexing by la tent semantic analysis. *Journal of the American Society of Information science*, vol. (416/6) : 391-407.

Fellbaum C. (Ed.) (1998). *WordNet : An Electronic Lexical Database*. MIT Press.

Hearst M.A. (1998). Automated discovery of Wordnet relations, In Fellbaum C. (Ed.), *WordNet : An Electronic Lexical Database*. MIT Press : 131-151.

Lafourcade M. et Prince V. (2001). *Relative Synonymy and Conceptual Vectors . NLPRS01* :127-134.

Larousse (1992), *Thésaurus Larousse - des idées aux mots - des mots aux idées.* , 1re édition (2e édition 2001).

Lee J.H., Kim M.H. et Lee Y.J. (1993). Information Retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, vol. (49/2) : 188-207.

Miller G.A. and Fellbaum C. (1991). Semantic Networks in English. In Levin B. et Pinker St. (Eds) *Lexical and Conceptual Semantics* : 197-229.

Palmer F.R. (1976). *Semantics : A New Introduction*. Cambridge University Press.

Resnik P. (1995). Using Information Contents to Evaluate Semantic Similarity in a Taxonomy. In *IJCAI-*

95.

Resnik P. (1999a). Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, vol. (11) : 95-130.

Resnik P. (1999b). Disambiguating noun groupings with respect to WordNet senses. In Armstrong S., Church K., Isabelle P., Tzoukermann E., Manzi S. and Yarouwsky D. (Eds.), *Natural Language Processing using Large Corpora*. Kluwer Academic.

Roget P. (1852). *Thesaurus of English Words and Phrases*. Longman.

Salton G. (1968). *Automatic Information Organisation and Retrieval*, McGraw-Hill.

Salton G. et MacGill M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Sparck J.K. (1986). *Synonymy and Semantic Classification*. Edinburgh Information Technology Serie.

Schwab D., Lafourcade M. et Prince V. (2002). *Antonymy and Conceptual Vectors.* In *Proceedings of* COLING'02, vol. (2/2).

Thelen M. et Riloff E. (2002). A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In *Proceedings of* EMNLP 2002, vol. (1/1).

Yarowsky D. (1992). *Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora*, *COLING'92* : 454-460.

Phylogenetic analysis of Gregory of Nazianzus' Homily 27

Anne-Catherine Lantin[1], Philippe V. Baret[1], Caroline Macé[2]

[1]Université catholique de Louvain – AGRO – GENA – 1348 Louvain-la-Neuve – Belgique
[2]Katholieke Universiteit Leuven – HIW – 3000 Leuven – België
lantin@gena.ucl.ac.be

Abstract

For a few years, biology and philology have come together by sharing a tool, namely the phylogenetic methods. Initially developed to establish relations between living organisms, these methods can also help reconstitute the history of multiple copies of an ancient text, which accumulated variations as time went by. This was already materialized by a few fruitful collaborations between philologists and biologists.

In this paper, the phylogenetic method of maximum parsimony was applied to 105 manuscripts containing the text of Gregory of Nazianzus' Homily 27. We aimed to compare results obtained by this phylogenetic analysis with what is known by philologists about the history of the text. The tree obtained agrees with philological evidences, but some undetermination remains for a few manuscripts.

This experiment brings new support to the pertinence of applying phylogeny approaches to manuscript traditions, which constitute a valid and powerful tool of classification.

Keywords: phylogeny, stemmatology, classification of manuscripts

1. Introduction

Phylogeny refers to the evolutive history of biological species, hypothetically reconstituted and represented as a tree. Stemmatology, a discipline from the field of philology, also builds trees, or stemmata, that tell how an unique ancient text gave birth to multiple different copies, as a result of variations introduced by scribes in the course of time. Both disciplines have thus in common that they try to infer the evolutionary history of their objects.

1.1. *Methodological and conceptual similarities*

In both disciplines, the scientific investigation can be seen as a three-step process. First, observing and comparing objects (living organisms or manuscripts) in order to build a dataset. Second, classifying the objects into groups according to their similarities. Third, if the data required is available, finding their origin and infering their evolutive history.

Objects can be analyzed only if they are reduced to a certain amount of characteristics. For living organisms, these characteristics consist in morphological, physiological, or molecular features (sequences of genes and proteins) and are called *characters*. For manuscripts, these features consist in all the positions within the text which vary from one reading to another; they are called *variant readings*.

Transmission and *variation* are key concepts in evolution, for organisms as well as for manuscripts.

Transmission of characters follows the basis hypothesis that an object inherits all its characteristics from the same ancestor: this is known as the unique ascendance hypothesis. However, this assumption is sometimes violated. In philology, it occurs when one manuscript is copied from two different templates, and this phenomenon is called *contamination*. In biology, it occurs when two species cross to give birth to a new one (*hybridation*) or when genetic material is transferred from one species to another (this phenomenon is known as *horizontal gene transfer*).

Characters are transmitted from one object to its descendants, but they also change in the course of time. In biology, *variation* is produced by genetic events (mutations); in philology, by alterations of the texts, introduced intentionally or unintentionally by those who copied them.

1.2. *Phylogenetic methods*

Phylogenetic methods are computerized algorithms that create trees, which graphically show relationships between objects. Different methods can be used and debate continues to center around the merits of competing methods of phylogeny inference. They can be applied to any set of data which contains objects subject to evolution (temporal changes), and as a result, they are also adapted to the construction of 'trees of manuscripts' (stemmata).

Extension of biological algorithms to philological applications was carried on over several manuscript traditions, giving consistent results (Mooney *et al*., 2003 ; Howe *et al*., 2001 ; Barbrook et *al*., 1998 ; Robinson and O'Hara, 1993).

The objective of the present paper is to use biological approaches to build a stemma of a Homily written by Saint Gregory of Nazianzus (ca. AD 330-390), one of the Cappadocian Church Fathers[1] and a bishop of Constantinople. The stemma aims to represent the evolution of Gregory's text from the 5th century (when the text was translated into Latin) to 1550 (when first printed versions of the text started to circulate).

2. Material and methods

2.1. *Dataset*

The dataset consists in a matrix of 141 lines and 556 columns. Lines correspond to all the extant manuscripts that contain Gregory of Nazianzus' Homily 27, and columns indicate all the positions within the text where the reading is different from one manuscript to another. As more than two variant readings can be found for the same position, the coding is not binary. For instance, entries in a matrix column can go from 0 to 7 when there are eight different variant readings for the same position. These codes do not reflect any chronological order, but '0' means that the manuscript has the same reading as the base text, an arbitrarily chosen manuscript used as a fixed comparison point. In our case, the base text is an edition from the 18th century, not included in the 141 manuscripts we studied.

The 141 lines correspond to 135 Greek manuscripts, 1 printed edition (EA), 1 artificial witness (Tanc) and 4 translations. Two translations are in Syriac (Tsyr1 and Tsyr2), one is in Latin (Tlat), and one is in Armenian (Tarm). We left aside 36 of the Greek manuscripts: 25 manuscripts are uninformative for editing the text, 8 manuscripts are too incomplete, and no microfilm is available for 3 manuscripts.

[1] The title of "Church Father" is given to the most important Greek and Latin authors from the 2nd to the 6th century, whose writings are considered founding texts of the tradition of the Church.

For these 105 witnesses, 86 out of 556 characters are constant and 282 variable characters are parsimony-uninformative, which means that they do not bring any information to build a phylogenetic tree under the parsimony principle. As a result, only 188 characters were taken into account for the analysis.

2.2. *Phylogenetic analysis*

The phylogenetic analysis was performed using PAUP* version 4.0b10 (Swofford, 1998). It is the result of a four-step process.

2.2.1. *Choosing the optimality criterion of parsimony*

Phylogenetic methods search for the tree that meets an optimality criterion by evaluating all possible individual trees generated out of the data. The optimality criterion called parsimony is based on the assumption that the most likely tree is the one that requires the fewest number of changes to explain the data. The basic premise of parsimony is that objects sharing a common characteristic do so because they inherited that characteristic from a common ancestor (in our case, that manuscripts sharing the same variant reading do so because they were copied from the same template). However, conflicts with that assumption often occur, because sharing a common characteristic may be due to independent parallel evolutionary events. For example, two scribes may independently introduce the same mistake in their copies. In biology, these independent evolutionary changes are gathered together under the term *homoplasy*. Homoplasies are regarded as 'extra' steps or hypotheses that are required to explain the data. Parsimony operates by selecting the tree or trees that minimize the number of evolutionary steps, including homoplasies. The number of steps required to fit the data with a given tree is referred to as the score of the tree.

Other optimality criteria exist, such as Maximum Likelihood, but they require precise assumption of the mechanisms of change and precise parameters of evolution; no such model exists for the evolution of manuscripts.

2.2.2. *Conducting a heuristic search*

When the number of manuscripts is small, it is possible to evaluate each of the possible trees. This exhaustive search guarantees finding the best tree (the most parsimonious one). But the number of possible trees rapidly increases with the number of manuscripts and an exhaustive search becomes impractical even for the most powerful computers (for ten manuscripts, there are more than 34 millions possible unrooted trees). When the number of trees is large, a heuristic strategy is used. A heuristic approach is essentially an algorithm in which an initial tree is selected, then rearrangements are sought to improve the tree. There are numerous algorithms of rearrangements. The one used in this experiment is the TBR search (tree bisection-reconnection): a provisional tree is cut into two subtrees at a branch, and these two subtrees are then reconnected by joining two branches, one from each subtree, to generate a different tree. This is tried for all possible pairs of branches of the two subtrees to generate many different trees.

2.2.3. *Rooting the trees*

Rooting a tree means choosing a point on the unrooted tree as representing the earliest time in the evolutionary history of the objects. An unrooted tree only represents groupings between the objects; a rooted tree shows a sequence of events, starting from the root. Rooting a tree can be done by designating a point in the middle of a branch as the common origin, or by selecting any one of the objects as an outgroup (an object designated as an outsider to the

other objects). In outgroup rooting, the point where the outgroup meets the ingroup is the ancestor of the whole ingroup. In this experiment, we rooted all the equally parsimonious trees with "Tanc". This witness was choosen as an ougroup because it is an artificially made manuscript, built with variant readings from Tarm and Tlat considered to be more ancient than the others.

2.2.4. *Computing the consensus tree*

Maximum parsimony methods often produce several parsimonious trees. In our analysis, more than 400 000 trees with the same score were generated. To make sense of this huge amount of equivalent trees, we built a consensus tree which sums up the common patterns observed in all the most parsimonious trees. In a strict consensus tree, any conflicting branching patterns for a set of objects among some rival trees are not resolved and result in a multifurcating branching pattern. This is why in Figure 1, some manuscripts are not organized in clusters but all have the same stauts according to the vertical branch they originate from.

3. Results

3.1. *Tree of the complete tradition*

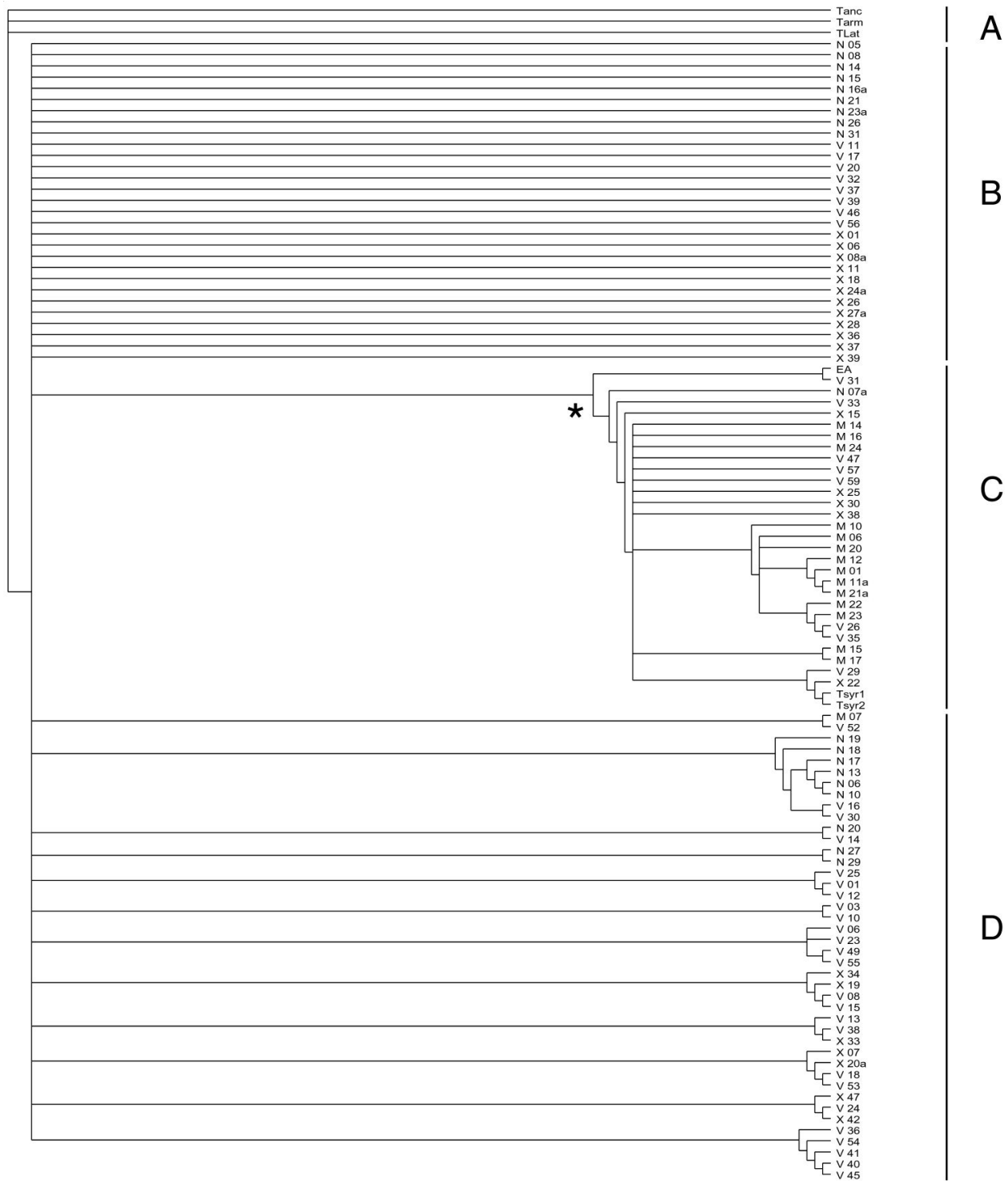


Figure 1. Consensus tree of the complete manuscripts tradition, created by PAUP and visualized with TREEVIEW. Codes on the right refer to the main clusters.*

3.2. *Interpreting the tree*

3.2.1. *The point of view of the biologist*

The tree is rooted with Tanc, which clusters with two translations, Tlat and Tarm [A].

Three other groups can be defined:

Group [B] contains 29 manuscripts, all deriving from the same vertical branch. This does not mean that all the manuscripts were copied from the same template, but rather that the relationships between these manuscripts are undefined, due to a lack of discriminant information.

Group [C] contains 31 clearly differentiated manuscripts. Within this group, clear substructures such as [Cm] and [Cs] are well defined (Figure 2).

Group [D] contains 42 manuscripts, clustered in small independent groups of 2 to 8 manuscripts. All manuscripts from each group should share a common ancestor, but no higher structure is observed.

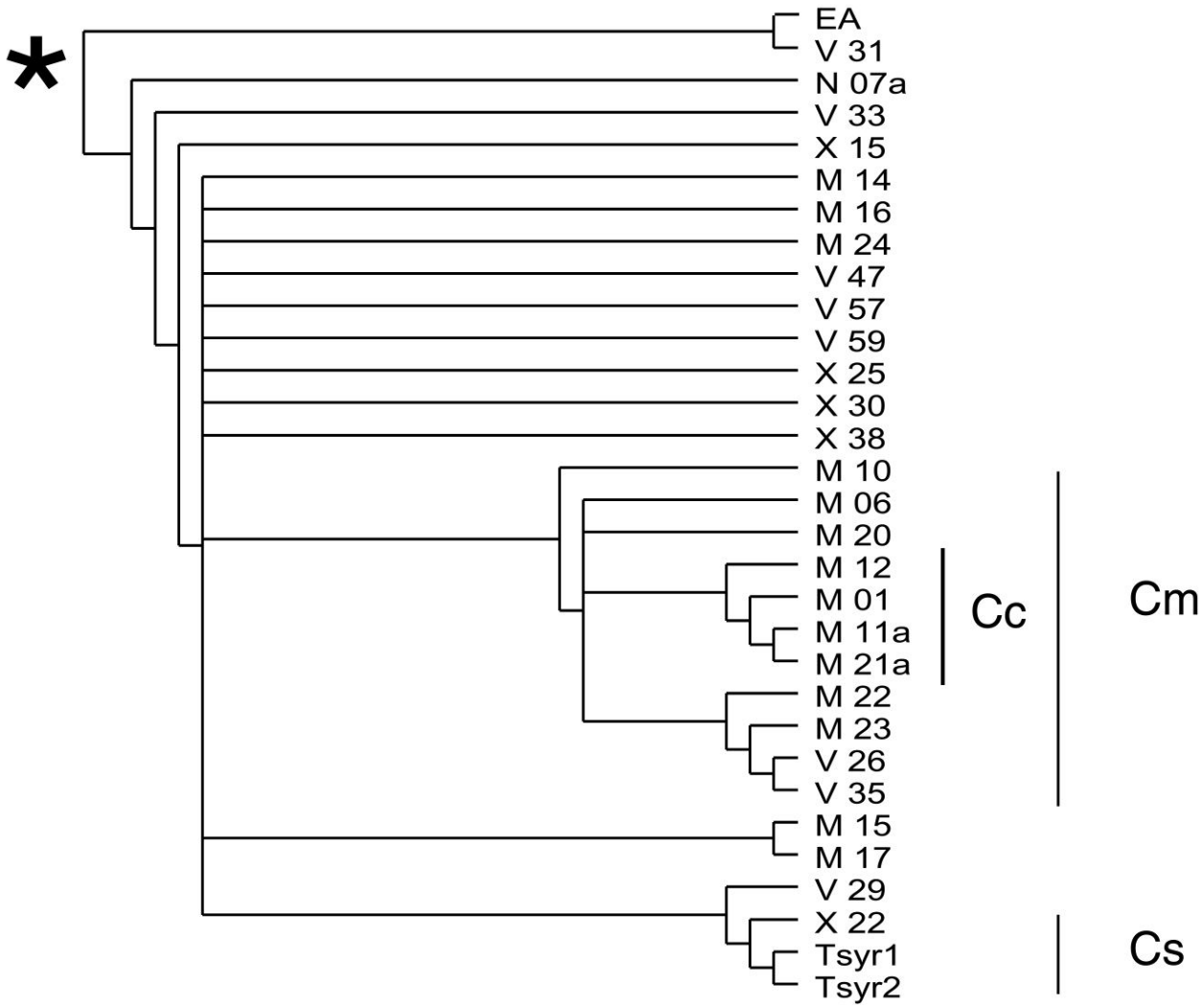


Figure 2. Details of the consensus tree of figure 1: the C group.

3.2.2. *The point of view of the philologist*

Reading the tree in Figure 1, the philologist can draw some conclusions about each one of the groups.

Group [A] clusters Tarm and Tlat, close to the root Tanc. As Tanc is a consensus made from Tarm and Tlat, it is not surprising to find them together in group [A]. It is also interesting to note that the two Syriac translations do not appear in this group, even if in Tsyr1 and Tsyr2 most data are missing, exaclty like in Tarm and Tlat. The presence of missing data is due to the way the translations were processed and discrepancies coded between Greek standard text and translations. All other statistical or computerized methods that were applied precedently (Baret, 2004; Macé, 2003) were unable to correctly handle missing data: Tsyr1 and Tsyr2 were always misclassified and misplaced, next to Tarm and Tlat, with which they have in common only missing data.

Manuscripts from group [B] are indeed difficult to classify, and this undetermination represents fairly well the problems faced by philologists with these manuscripts, due to the absence of discriminant variant readings.

Most of the manuscripts in the [Cm] group are from Southern Italian origin, as opposed to the other manuscripts. Interestingly enough, the Syriac translations (Tsyr1 and Tsyr2) [Cs] belong to the same main group [C] as this Southern Italian sub-group [Cm]. It is indeed well known that strong cultural connections existed between Syria and Southern Italy in the Byzantine world before the time when our oldest preserved Greek manuscripts were written. It is thus plausible that Tsyr1 and Tsyr2 were translated from the same template as the one used for the Greek copies from the [Cm] group. Paleographical evidence also shows that all manuscripts from group [Cc] were copied by Calabrian monks. It is thus exact to group them is the same cluster.

Two subgroups from group [D] are also corroborated by external philological criteria. The first group contains manuscripts N.18, N.17, N.13, N.06, N.10, V.16 and V.30, which share the same order of homilies and the second group contains manuscripts V.36, V.54, V.41, V.40 and V.45, which share the same 'mise en page' and marginal notes.

4. Conclusion

By applying the maximum parsimony method to the manuscripts containing Homily 27, we generated a tree which is validated by philological evidence and which offers new hypotheses that sould be investigated by philologists. The most parsimonious tree could not be found, due to limits in computing time, so the tree presented is a consensus of more than 400 000 equally parsimonious trees. In this tree, the two Syriac translations were correctly handled, which was not the case when other statistical or computerized methods were used.

From a technical point of view, phylogenetic methods can be easily transferred to philological data. The tree of Figure 1 is the first complete tree ever drawn showing relationships between all manuscripts of an Homily of Gregory of Nazianzus, although research on this complicated manuscript tradition has been carried on for more than one hundred years. Some of the groups are corroborated by external evidences (same origin, similarities of scripture or 'mise en page') and despite the fact that some relationships remain undetermined, the structure of the tree broadly agrees with what is know about the text.

A conceptual analysis reveals that some problems of phylogenetic inference in biology also apply in philology. For instance, *hybridation* and *horizontal gene transfer* correspond to the philological phenomenon of *contamination*; *homoplasies* also exist for manuscripts; and *rooting the trees* requires in biology as well as in philology some information over the age of manuscripts or the chronology of variant readings. *Contamination* and *homoplasises* were not

taken into account in this analysis, but identifying ambiguous manuscripts and characters and removing them from the dataset could help reduce the number of equally parsimonious trees.

Although phylogenetic inference still has to face technical and conceptual challenges, we are conviced that it is the most powerful tool offered to the philologist in order to explore a manuscript tradition.

PAUP procedure

```
execute nazianze-complet.nex
taxset exclus = 2-26 31 39 75 86 87 133 28 116 126 71 64
delete exclus / only
set criterion = parsimony
charset noninform = uninf
exclude noninform / only
hsearch addseq = simple reftax = 65
Select 'Reset Maxtrees'
Select 'automatically increase by 200'
outgroup Tanc
roottrees
savetrees file = nazianze.tre
contree/ treefile = consensus.tre
```

Acknowledgements

We would like to thank all the members of the Center for the Study of Gregory of Nazianzus (http://nazianzos.fltr.ucl.ac.be), and particularly Marc Dubuisson who built all the databases. We would also like to thank A. Vanderpoorten for his helpful comments on the computer analysis.

References

Barbrook A.C. *et al.* (1998). The phylogeny of the *Canterbury Tales*. *Nature*, vol. (394): 839.

Baret P.V. *et al.* (accepted). Experimental phylogenetic analysis of a Greek manuscript tradition. In *Proceedings of the Washington Academy of Sciences*.

Hall B.G. (2001). *Phylogenetic Trees Made Easy*. Sinauer Associates, Inc.

Howe C.J. *et al.* (2001). Manuscript evolution. *Trends in Genetics*, vol. (17): 147-152.

Lantin A.-C. (2003). *Phylogénie des manuscrits de Grégoire de Nazianze.* Final dissertation. UCL.

Mooney L.R. *et al.* (2003). Stemmatic Analysis of Lydgate's Kings of England: a test case for the application of software developed for evolutionary biology to manuscript stemmatics. *Revue d'Histoire des Textes*, vol. (31): 202-240.

Nei M. and Kumar S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press.

Macé C. *et al.* (2003). Le classement des manuscrits par la statistique et la phylogénétique: le cas de Grégoire de Nazianze et de Basile le Minime. *Revue d'Histoire des Textes*, vol. (31): 241-273.

O'Hara R.J. and Robinson P.M.W. (1993). Computer-assisted methods of stemmatic analysis. *The Canterbury Tales Project Occasional Papers*, t. (I): 53-74.

Page R.D.M. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, vol. (12): 357-358.

Swofford D.L. (1998). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Validité des visualisations de données textuelles

Ludovic Lebart

CNRS - ENST – 46, rue Barrault – 75013 Paris
lebart@enst.fr

Abstract

Recent trends of research show that exploratory techniques (principal axes methods and unsupervised clustering as well) are widely used and have great potential in the processing of textual data. However, the problem of the validation of the obtained results is seldom addressed. We will focus our paper on the use of meta-data for external validation, and on the assessments of visualizations. The examples of application concern the statistical processing of responses to open-ended questions in the context of an international survey.

Résumé

Les méthodes de statistiques exploratoires (analyses en axes principaux, classification et cartes de Kohonen), contribuent à la description des données textuelles, description venant d'ailleurs en complément, et non en substitution, des approches plus manuelles ou traditionnelles. On peut regretter que dans la pratique, les visualisations obtenues soient rarement accompagnées de procédures de validation. On évoque ici les méthodes de *validation externe*, utilisant la meta-information disponible, mais on insiste surtout sur les méthodes de *validation interne* qui font intervenir les techniques de rééchantillonnage (principalement : *bootstrap*). Les exemples d'application concernent des réponses à une question ouverte dans une enquête internationale.

Mots-clés : lexicométrie, analyse de données textuelles, validation, *bootstrap*.

1. Introduction

Les méthodes d'analyse descriptive des données multidimensionnelles qui constituent le cœur de la méthodologie de l'analyse statistique exploratoire des données se perfectionnent en s'efforçant d'utiliser d'une part la méta-information disponible (informations que l'on possède sur le tableau de données que l'on analyse, et qui ne figurent pas dans le tableau lui-même) d'autre part les techniques de validation modernes essentiellement fondées sur les méthodes de rééchantillonnage (Monte-Carlo, validation croisée, Bootstrap). On peut schématiser ces récents développements en disant qu'ils tentent de répondre aux deux questions :

– *Comment utiliser ce que l'on sait pour essayer d'en savoir plus ?*

– *Comment valider ou généraliser les résultats obtenus ?*

2. La complexité des données textuelles : point de vue d'un statisticien

Les concepts de variables et d'observations sont beaucoup plus complexes que ceux de la statistique usuelle. Montrons-le en six points.

2.1. *Des variables construites*

Les variables ne sont pas définies *a priori*, mais dérivées à partir des textes. On peut considérer comme variables (qualitatives) les formes graphiques, les lemmes, les segments, mais des

seuils de fréquences interviendront qui modifieront le nombre et la nature de ces variables, pour un même recueil de textes.

2.2. *Des unités statistiques diverses*

Les observations, ou individus peuvent être des documents, des abstracts ou des brevets, des répondants (dans une enquête comportant des questions ouvertes, par exemple), des paragraphes, des phrases, des discours, des auteurs… et, comme nous le verrons plus bas, plusieurs niveaux peuvent coexister dans un même recueil, ce qui ne manquera pas de poser des problèmes lorsqu'il s'agira de procéder à des inférence statistiques.

2.3. *Des tableaux creux*

Les données numériques résultantes seront très volumineuses, clairsemées (*sparse*), et très bruitées. Ce dernier terme, emprunté à la théorie du signal, signifie en bref que les structures qui nous intéressent ne constitueront qu'une faible part de la variabilité (variance) totale.

2.4. *Une méta-information considérable*

Il existe un volume important de méta-information, c'est-à-dire d'informations externes aux tableaux analysés. Pour un statisticien, la meta-information disponible dans le cas de données textuelles est quelque chose de démesuré, si on compare cette meta-information à celle des recueils de données numériques ou qualitatives habituels (fichiers d'enquête, par exemple). Chaque mot utilisé, même si c'est un mot grammatical (mot vide, ou encore mot-outil) a droit à plusieurs lignes, voire plusieurs pages dans un dictionnaire…

2.5. *Des processus, et non des vecteurs*

Le recueil de données est globalement séquentiel (analyse d'un très grand texte) ou localement séquentiel (analyse de documents, de réponses). Prenons le cas de la réponse d'un individu i à une question ouverte. Si $\{ s_1, s_2, \ldots , s_v \}$ désigne un ensemble de v éléments différents (le vocabulaire), un individu statistique i ($i \leq n$) pourra être caractérisé par une séquence ordonnée de longueur variable $\gamma(i)$: $\{ s_{r(i,1)}, s_{r(i,2)}, .. , s_{r(i,\gamma(i))}\}$, où : $1 \leq r(i,k) \leq v$, et $1 \leq k \leq \gamma(i)$. Un mot peut apparaître plusieurs fois dans une réponse. La quantité r(i, k) est alors l'index repérant le k-ième mot dans la réponse de l'individu i.

Calculer des similarités ou des distances entre séquences ordonnées de longueurs variables avec répétitions n'est pas une tâche courante de la statistique exploratoire multidimensionnelle. L'introduction d'unités composites, comme les segments, permet une prise en compte partielle de l'ordre des mots dans les phrases. On peut aussi utiliser les informations données par un arbre syntaxique (dont l'obtention ne peut être entièrement automatique) dans les calculs de similarités entre phrases ou réponses. Notons que des procédures d'analyse statistique multidimensionnelle, comme l'analyse des correspondances d'une table de contingence croisant des mots (en ligne) et des parties de textes (en colonne) ne font appels qu'à des modèles statistiques rudimentaires, mais peu contraignant. La méthode est une sorte de bilan global, qui ne prend pas en compte la nature séquentielle du texte, mais qui peut intégrer toutes les nouvelles variables construites après une analyse linguistique.

2.6. *Particularités statistiques des réponses aux questions ouvertes*

Dans les réponses aux questions ouvertes, le statut de la fréquence des mots est ambigu. Les fréquences lexicales observées sont pour une large partie artificielles, car la même question est parfois posée à des centaines ou des milliers de personnes... la juxtaposition des réponses constitue un « texte » redondant par construction. Il existe en fait deux niveaux d'individus

statistiques : les personnes interrogées, qui sont les individus habituels des statisticiens d'enquêtes, et les mots utilisés, qui sont, les individus les plus familiers pour les linguistes quantitatifs. Certains tests statistiques seront significatifs au niveau des comptages de mots, mais pas au niveau des comptages d'individus, et donc ne permettront pas une inférence des résultats à la population interrogée. Les analyses de réponses regroupées seront assez voisines des analyses de « vrais textes » (littéraires, politiques, historiques), alors que les analyses directes de réponses seront voisines des traitements effectués en recherche documentaire.

3. Les analyses statistiques ; les outils de base

3.1. *Analyse des correspondances et classification*

Les analyses des correspondances (A.C.) et les méthodes de classification peuvent décrire les tables de contingence croisant les réponses et les formes graphiques (Benzécri, 1981), ou des groupes de réponses (par exemple regroupement selon le niveau d'instruction des répondants) et les formes graphiques. Elles permettent de visualiser sous forme de séries de cartes planes (ou de dendrogrammes, ou de cartes auto-associatives de Kohonen) les associations entre mots (formes) et groupes ou modalités. Ainsi, une visualisation des proximités entre mots et catégories socioprofessionnelles pourra aider la lecture des réponses de chacune de ces catégories. Il est toujours intéressant de comparer les analyses sur textes bruts et sur textes lemmatisés, analyses qui, selon les cas, se confirment ou se complètent.

Toutes ces visualisations graphiques, parce qu'elles sont souvent suggestives, sont habituellement produites sans critère de validation, dans un premier temps faute de procédures disponibles, puis, probablement, par absence de demande pressante de la part des utilisateurs.

3.2. *Un exemple : visualisation simultanée mots-catégories par cartes de Kohonen*

Nous illustreront cette brève évocation des outils usuels par un exemple de visualisation à l'aide de cartes de Kohonen (Kohonen, 1989) qui, appliquées au système de distances de l'A.C., permettent de représenter simultanément les lignes et les colonnes de la table lexicale.

Contexte de l'étude : une question ouverte a été posée lors d'une enquête multinationale (cf. Hayashi *et al*., 1992) avec le libellé suivant "*Que signifie pour vous la culture de votre pays?*". C'est le sous-échantillon relatif à la France qui est traité ici (effectif : 1009 personnes représentatives des personnes de 15 ans ou plus). L'ensemble des réponses à cette question représente 14742 occurrences de 2248 mots (formes graphiques) distincts. Sont traités ici les 111 mots apparaissant au moins 20 fois, qui représentent quand même 10005 occurrences, auxquels on a retranché 40 mots-outils ou mots vides usuels, ce qui porte à 71 le nombre définitif de mots gardés. Les réponses sont ici regroupées en neuf catégories obtenues par croisement de l'âge en trois classes (moins de 30 ans, 30-55 ans, plus de 55 ans) et du niveau d'éducation en trois classes (bas, moyen, haut).

La carte de Kohonen utilise ici le même système de distance que celui de l'analyse des correspondances de la table lexicale qui croise les occurrences des 71 mots les plus fréquents et les neuf catégories précédentes (les identificateurs des catégories sont abrégés). On obtient ainsi une synthèse assez comparable à celle de l'analyse des correspondances qui figure plus loin (figure 2). Cette synthèse nous permet de repérer assez rapidement des cooccurrences de mots dans une même catégorie, et des proximités lexicales entre catégories. Les premiers avantages de la représentation de la figure 1 dans le cas textuel concernent le format d'édition, composé de fenêtres rectangulaires particulièrement aptes à recevoir des listes de mots.

littérature grands art +55/HAUT	niveau esprit comme -30/HAUT	france	écrivains surtout histoire cinéma	peut grand	patrimoine
vivre faire arts 30-55/HA		langue	français chose		notre nos
tous musique	pense aussi	peu pays par française culture		enfants autres	même 30-55/MO
être trop gens +55/MOYE	très ont fait	monde	temps	beaucoup	mais
peinture	vie faut	sont nous bien	pour plus jeunes ils +55/BAS	pas	livres
théâtre quand moins cultivés -30/BAS	choses assez 30-55/BA	mal façon bonne		école télévision sais	sait rien lecture -30/MOYE

Figure 1. Représentation par carte de Kohonen des proximités entre mots et entre catégories

S'agissant d'une classification (avec contraintes, pour les classes, de figurer dans une grille), on peut penser que les regroupements prennent en compte des informations qui vont au-delà de celles données par les deux dimensions d'un premier plan factoriel.

Parmi les inconvénients de cette technique de visualisation, citons le caractère « non déterministe » (comme les méthodes de classification autour de centres mobiles ou *k-means* usuelles) des algorithmes qui produisent, dans ce cas particulier, des cartes différentes en fonction de l'initialisation aléatoire de départ. Enfin, il n'existe pas vraiment, dans les implémentations disponibles, de méthodes de validation permettant de décrire la confiance à accorder à la fois à la composition et aux positions relatives des différentes cases de la grille.

La seule validation possible, commune avec les méthodes de classification du type *k-means* ou nuées dynamiques, consiste à opérer divers démarrages aléatoires de l'algorithme, et à observer visuellement la stabilité des résultats. De simples clics de souris génèrent instantanément de nouvelles cartes, qui peuvent différer à la fois par les regroupements à l'intérieur des cellules et par les positions relatives des cellules.

4. La validation externe pour les analyses « en axes principaux »

Le principe de la validation externe consiste à confronter le corpus, ainsi que toute structure déduite exclusivement du corpus, avec des informations externes concernant des composantes du corpus. Un exemple courant est fourni par, d'une part un corpus de réponses à une question ouverte, et d'autre part un tableau des caractéristiques des répondants à la question.

4.1. *Les valeurs-test*

L'information externe se manifeste souvent sous forme de « variables supplémentaires » (ou « variables illustratives » qui sont positionnées (en fait : projetées) *a posteriori* sur les axes principaux. Pour chaque axe et chaque variable, une « valeur-test » exprime la signification statistique de la coordonnée de la variable sur l'axe (sous l'hypothèse d'indépendance de la variable et des variables actives de l'analyse).

Supposons qu'une modalité supplémentaire j concerne k_j individus (répondants d'une enquête, documents, etc.). Si ces k_j individus sont tirés au hasard (hypothèse nulle H_0) parmi les k individus analysés (tirage supposé sans remise), la moyenne de k_j coordonnées $\psi_{\alpha i}$ tirées au hasard dans l'ensemble fini des k valeurs est une variable aléatoire $X_{\alpha j}$, avec pour espérance (moyenne) :E $(X_{\alpha j})= 0$,(origine de l'axe α au point moyen) et pour écart-type : $v(j, k_j)$.

Le carré de cet écart-type vaut $v(j,k_j)^2 = \frac{1}{k_j}\left(\frac{k-1}{k-k_j}\right)\lambda_\alpha$ (*cf.*, *e.g.*, Lebart *et al.*, 1998).

Alors, $t_\alpha(j) = \varphi_{\alpha j} / v(j, k_j)$ est une variable centrée réduite ($\varphi_{\alpha j}$ est la coordonnées du point-modalité j sur l'axe α). La quantité $t_\alpha(j)$ mesure en nombre d'écart-types la distance entre la modalité j, c'est-à-dire le quasi-barycentre des k_j individus, et l'origine sur l'axe factoriel α. On appelle cette quantité « valeur-test ». On montre que sa distribution tend vers une loi normale centrée réduite lorsque k augmente indéfiniment. La valeur-test est un critère qui permet d'apprécier rapidement si une modalité a une position « significative » sur un axe. On considère comme occupant une « position significative » les modalités dont les valeurs-test sont supérieures à *2* en valeur absolue, correspondant approximativement au seuil *5 %*. On note que les valeurs-test n'ont de sens que pour les modalités supplémentaires.

4.2. *Le problème des comparaisons multiples*

Le calcul simultané de plusieurs valeurs-test ou de plusieurs seuils de probabilités se heurte à l'écueil des *comparaisons multiples,* bien connu des statisticiens. Supposons que l'on projette *100* modalités supplémentaires qui soient vraiment tirées au hasard. Les valeurs-test attachées à ces modalités sont alors toutes des réalisations de variables aléatoires normales centrées réduites indépendantes. Dans ces conditions, en moyenne, sur *100* valeurs-test calculées, *5* seront en dehors de l'intervalle [*-1.96* , *+1.96*], et *5* dépasseront la valeur *1.65* (test unilatéral). Le seuil de *5 %* n'a de sens en fait que pour un seul test, et non pour des tests multiples. On résout de façon pragmatique cette difficulté en choisissant un seuil plus sévère. Plusieurs solutions ont été proposées pour ce faire : la méthode de *Bonferroni* recommande de diviser le seuil statistique par le nombre de tests (ici, le nombre de valeurs-test calculées). Cette réduction du seuil est souvent considérée comme étant trop sévère (Hochberg, 1988 ; Perneger, 1998). Des études approfondies de ce problème se trouvent dans Hsu (1996), Westfall et Young (1993).

5. Validation interne : le re-échantillonnage.

Les visualisations provenant des analyses en axes principaux [décomposition aux valeurs singulières (SVD), composantes principales (ACP), correspondances (AC)] n'ont de sens que si elles sont accompagnées de la confiance que l'on peut accorder à la position de chaque point. Or ces visualisations sont d'une grande complexité analytique, et il est exclu de procéder à des évaluations selon les méthodes de la statistique classique.

5.1. *Principe des zones de confiance "bootstrap"*

La technique de *bootstrap* (cf. Efron et Tibshirani, 1993) va permettre de tracer des zones de confiance (ellipses ou enveloppes convexes de réplications) autour des points représentés sur les plans principaux, que ces points représentent des mots ou des textes. La méthode consiste à construire *n* "réplications" de l'échantillon par tirage *avec remise* des unités statistiques. Dans une réplication, certaines unités apparaîtront ainsi deux fois ou plus, d'autres n'apparaîtront pas. On crée ainsi une variabilité autour du tableau de données de départ. Sur chacune des *n* réplications, on peut faire les calculs les plus complexes, comme ceux que nécessitent les techniques de visualisation. Sous des hypothèses faibles, on montre que la variabilité observée sur les *n* réplications est de l'ordre de grandeur de celle que l'on aurait observée dans la population. Autrement dit, on peut disposer de *n* réplications de paramètres complexes, et donc obtenir des intervalles de confiance pour ces paramètres. Cette propriété est valable pour les vecteurs propres, les réplications des valeurs propres étant biaisées.

5.2. *Cas des composantes principales et de la SVD*

Dans le cas des composantes principales, on relève plusieurs variantes de la méthode : le *bootstrap total* consiste à refaire une analyse complète pour chaque réplication. Cette procédure engendre une difficulté : les axes répliqués ne sont pas forcément homologues d'un réplication à une autre, il peut y avoir des interversions d'axes, voir des rotations. Il faut alors faire coïncider par des techniques d'*analyses procustéennes* les axes homologues (Markus, 1994 ; Milan et Whittaker, 1995). Le *bootstrap partiel* permet de lever cette difficulté. Il part de la constatation que le tableau initial est plus proche de la réalité observée que tous les tableaux répliqués, qui en sont des perturbations. L'analyse et les plans principaux de ce tableau initial servent de référence pour la projection de tous les tableaux répliqués (lignes et colonnes) en tant qu'éléments supplémentaires. Des expérience intensives (cf. Lebart *et al.*, 2003) ont montré l'efficacité de cette méthode.

5.3. *Cas de l'analyse des correspondances*

Les travaux de pionniers furent ceux de Gifi (1981), Greenacre (1984). Dans le cas d'une table lexicale, la technique consiste à tirer avec remise les occurrences de formes retenues. Ce tirage se fait selon un schéma multinomial comportant autant de catégories que la table a de cellules, et dont les fréquences théoriques sont celles des cellules. Les lignes et colonnes des tables répliquées sont alors projetées comme éléments supplémentaires sur les plans factoriels de l'analyse de la vraie table lexicale (*bootstrap partiel*). Des ACP des « nuages de réplications » correspondant à chaque point-élément (ligne ou colonne) fournissent les ellipses de confiance cherchées. Dans le cas d'un *bootstrap total*, on retrouve les difficultés déjà mentionnées pour les valeurs propres (voir, par exemple, Alvarez *et al*., 2002).

5.4. *Un exemple de zones de confiance*

La figure 2 représente de telles zones de confiances dans le premier plan factoriel de l'AC de la table lexicale (déjà utilisée section 3.2) qui croise les occurrences des 71 mots et les 9 catégories précédentes. Comme il n'est pas possible de représenter sur un même graphique toutes les ellipses, on a choisi quatre catégories (encadrées) en excluant les âges et les niveaux d'éducation moyens. On a de même choisi cinq mots (formes graphiques, en fait) de façon à recouvrir les différentes zones du graphique *(arts, chose, choses, cinéma, télévision*). Remarquons que les zones de confiance des catégories sont relativement petites, indiquant des positions (et des positions relatives) assez significatives pour ces catégories (la plus grande concerne les jeunes peu instruits, catégorie dont l'effectif est le plus faible (13 réponses dans l'échantillon). On voit qu'à la mention du mot *télévision* correspond une longue ellipse de

confiance, qui fait osciller la position du point entre les catégories –30/BAS (jeunes peu instruits) et +55/BAS (seniors peu instruits) mais reste du côté de ces classes de niveaux d'éducation. On note le comportement inverse du mot *arts*. Enfin, on a représenté les positions de deux formes graphiques (*chose* et *choses*) du même lemme *chose*, pour montrer qu'il peut arriver que les positions de telles formes soient significativement distinctes (les ellipses ne se recouvrent pas) (il semble que *chose* au singulier soit souvent associé à la réponse *pas grand chose*, qui exprime une opinion critique ou désabusée vis-à-vis de la culture). Comme souvent, la forme graphique brute contient des informations de contexte. Il est intéressant de noter que l'augmentation du nombre de mots consécutive au choix d'un seuil plus bas a en général pour effet de diminuer la taille des ellipses relatives aux catégories, malgré la prise en compte de données évidemment plus bruitées (mots plus rares).

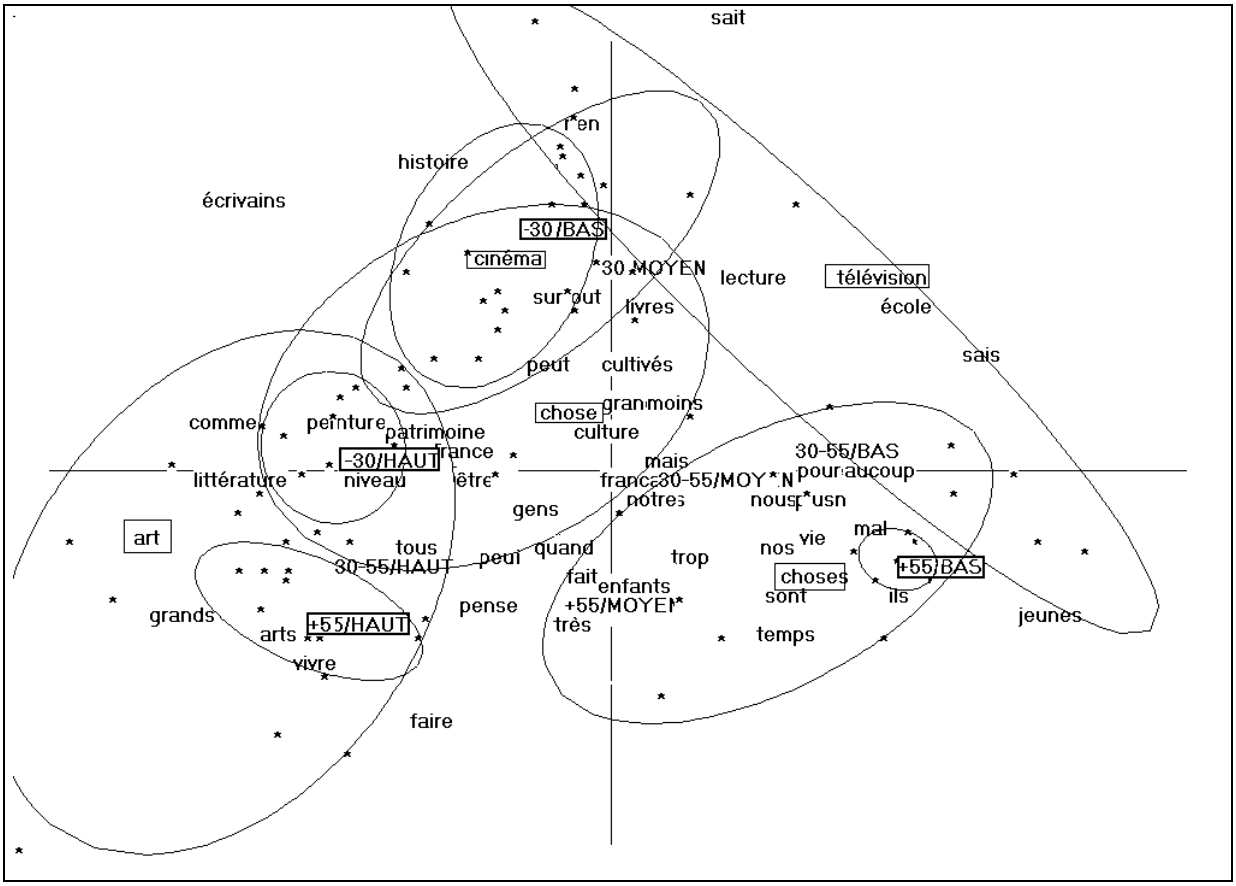


Figure 2. Zones de confiance pour 4 catégories de répondants et pour 5 mots.

5.5. *Les différentes unités statistiques*

Nous avons vu, section 2.2, qu'il existait plusieurs niveaux d'unités statistiques. Un exemple est donné par les questions ouvertes : l'individu interrogé constitue l'unité classique des statisticiens, alors que l'occurrence des mots est celle des lexicomètres (cf. Tuzzi et Tweedie, 2000). La souplesse de la procédure de bootstrap permet d'adapter les réplications aux deux préoccupations (tirage avec remise des répondants, ou des occurrences de mots), et obtenir ainsi des zones de confiance adaptées aux deux hypothèses de départ. On retrouve la même situation en *Web-log mining*, (mots, pages, sites) ou encore en recherche documentaire (mots, titres, abstracts, textes complets).

6. Conclusion

Jusqu'à une date récente, les matériaux textuels, complexes, volumineux, riches en informations externes, étaient mal adaptés aux formalisations et aux épreuves de validité de la statistique classique. Les techniques de ré-échantillonnage fournissent un outil de validation souple, permettant de transformer une visualisation plaisante en un document scientifique.

[Les calculs et figures de cet article ont été réalisés avec le logiciel académique DTM]

References

Alvarez R., Bécue M., Lanero J.J. et Valencia O. (2002). Results stability in Textual Analysis: its Application to the Study of the Spanish Investiture Speeches (1979-2000). In *Actes des JADT 2002* : 1-12.

Benzécri J.-P. et collaborateurs (1981). *Pratique de l'analyse des données,* t. (3). Linguistique & Lexicologie. Dunod.

Efron B. et Tibshirani R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Efron B. (1979). Bootstraps methods : another look at the Jackknife. *Ann. Statist*., vol. (7) : 1-26.

Gifi A. (1981). *Non Linear Multivariate Analysis*. Department of Data theory, University of Leiden. (Nouvelle version, même titre : Wiley J., Chichester, 1990).

Greenacre M. (1984). *Theory and Applications of Correspondence Analysis*. Academic press.

Hayashi C., Suzuki T. et Sasaki M. (1992). *Data Analysis for Social Comparative Research: International Perspective*. North-Holland.

Hochberg Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, vol. (75) : 800-803.

Hsu J.C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall.

Kohonen T. (1989). *Self-Organization and Associative Memory*. Springer Verlag.

Lebart L., Piron M. et Steiner J.-F. (2003). *La Sémiométrie*. Dunod.

Lebart L., Salem A. et Berry E. (1998). *Exploring Textual Data*. Kluwer Ac. Publisher.

Markus M.Th. (1994). Bootstrap Confidence Regions for Homogeneity Analysis; the Influence of Rotation on Coverage Percentages. In Dutter R. et Grossmann W. (Eds), *COMPSTAT 1994*. Physica Verlag : 337-342.

Milan L. et Whittaker J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Appl. Statist*., vol. (44) : 131-149.

Perneger T.V. (1998). What is wrong with Bonferroni adjustments. *British Medical Journal*, vol. (136) : 1236-1238.

Tuzzi A. et Tweedie F.J. (2000). The best of both worlds: Comparing Mocar and Mcdisp. In *Actes des JADT 2000* : 271-276.

Westfall P.H. et Young S.S. (1993*). Resampling Based Multiple Testing: Examples and Methods for p-values Adjustment.* Wiley.

Autour du *Je* présidentiel

Jean-Marc Leblanc, Pierre Fiala

CEDITEC – Université Paris12 – Val de Marne – France
leblanc.jeanmarc@free.fr, fiala@univ-paris12.fr

Abstract

The functional comparison of various approaches to cooccurrence in some robust standard textual analyzers makes it possible to develop an experimental method in the courses of textual data processing intended for researchers in social sciences, which is too often absent from the field. Using a corpus of highly ritualized institutional speeches (New year addresses of the presidents of the French 5th Republic, 1958-2003), labelled or not, various functions to measure cooccurrence with LEXICO3, WEBLEX, HYPERBASE, ALCESTE are used in turn and compared. Different cooccurrence data are thus highlighted through figures, sorting, and graphic display, each leading to specific interpretations. The comparison shows the absolute necessity of in-depth experimentation in the analysis of textual data before any hermeneutic venture.

Keywords: cooccurrences, lexicogrammes, labeling, political discourse.

Résumé

La comparaison fonctionnelle de diverses approches des cooccurrences dans quelques analyseurs textuels standard robustes permet de développer dans les formations d'informatique textuelle destinées aux chercheurs en sciences humaines, une démarche expérimentale, qui fait trop souvent défaut dans le domaine. À partir d'un corpus de discours institutionnels fortement ritualisés (messages de vœux des présidents de la 5e république, 1959-2001), étiqueté ou non, diverses fonctions cooccurrentielles de LEXICO3, WEBLEX, HYPERBASE, ALCESTE sont successivement utilisées et comparées. Elles mettent en évidence, par les calculs, les tris, et les présentations graphiques des faits de cooccurrences différents, susceptibles d'interprétations complémentaires. La comparaison permet d'insister sur la nécessité absolue d'une expérimentation approfondie dans le traitement des données textuelles avant toute entreprise herméneutique.

Mots-clés : cooccurrences, lexicogrammes, étiquetages, discours politique.

1. Développer l'expérimentation et la comparaison en analyse textuelle

Trente ans après la parution des premiers travaux représentatifs du domaine, l'analyse des données textuelles poursuit en France son développement comme champ interdisciplinaire où se côtoient des disciplines diverses qui confrontent leurs hypothèses et leurs méthodes : statistique, linguistique, sociologie, histoire, communication ou philosophie politiques, histoire littéraire, analyse documentaire. Des ouvrages récents témoignent de la permanence du débat (Chateauraynaud, 2003), quand ce n'est pas des polémiques (Labbé, 2003), rendues malheureusement stériles parce que plus attachées à des promotions disciplinaires ou personnelles qu'au développement du domaine. Les progrès des interfaces permettent heureusement maintenant de comparer utilement les outils et les développements qui sont proposés pour traiter les problèmes.

Le renouvellement de l'intérêt pour l'analyse du discours et de la communication politiques, stimulé par les recherches sur l'argumentation et les mécanismes de persuasion, mais aussi les controverses sur l'utilisation des données quantitatives, ainsi que les utilisations sauvages des outils statistiques dans les débats politico-médiatiques nous ont incités à développer depuis quelques années une démarche de formation à la recherche où nous soulignons l'importance d'une plus large expérimentation. Pourquoi et comment comparer les textes ? Comment tester et optimiser des outils logiciels toujours plus nombreux, analogues sans être pourtant identiques, ni dans leur esprit ni dans leurs fonctionnalités ni dans leur résultat.

Le calcul des cooccurrences, ou mesure probabilisée des attirances entre formes dans un contexte donné, qui repose sur une notion bien décrite dans le domaine, a connu plusieurs types d'implémentations et de nombreuses exploitations ces dernières années. Mesure des cooccurrences binaires et recherche des lexicogrammes (Lafon, 1984 ; Tournier, 2003 ; Heiden, 2003), voisinage (Labbé, 1990), cooccurrences des formes spécifiques (Salem, Martinez, 2003). Cette problématique est propice à une démarche expérimentaliste visant à mettre en évidence les propriétés exploitables de chaque méthodologie.

2. *Je* dans le corpus des vœux présidentiels (1959-2001)

À l'instar de Labbé (1990) et Tournier (2003), nous examinons, le statut du *je* présidentiel, réduit ici à sa forme graphique, en explorant ses espaces cooccurrentiels dans les messages de vœux aux Français sous la Cinquième République (1959-2001). Nous mobilisons différents outils statistiques dans une démarche expérimentale, cumulative et comparée afin de mettre en lumière les réseaux lexicaux mais aussi sémantiques et thématiques qui gravitent autour de la marque personnelle centrale dans ce genre institutionnel. Les réseaux de cooccurrence, saisis par ces approches différentes apportent un éclairage sur la construction de l'ethos présidentiel au sein d'une forme fortement ritualisée mais aussi sur les outils eux-mêmes

2.1. *Les paramètres du corpus*

Les interventions produites de décembre 1959 à décembre 2001 soit 43 discours représentent un volume textuel de 41125 occurrences pour 5203 formes.

Ces discours forment une série textuelle chronologique, mais ils sont abordés ici dans leur dimension synchronique. Nous menons l'expérience sur le corpus dans son ensemble, tout en restant attentifs aux emplois individuels lorsqu'il s'agit de retourner au texte

2.2. *Les marques énonciatives*

Les spécificités par locuteur des pronoms personnels et adjectifs possessifs mettent en lumière le contraste des profils énonciatifs qui caractérise ce discours pourtant très ritualisé.

À l'énonciation fortement personnalisée, centrée sur le *je* de Giscard (*je*, +5), mais multipliant aussi les marques en direction des Français (*vous*, +20), on opposera en premier lieu l'importante prise de distance produite chez le général de Gaulle par le rejet des marques de la première personne du singulier (*je*, -16) et de la seconde du pluriel (*vous*, -25), la prise en charge de l'énoncé étant assurée par un *nous* dont le référent est essentiellement la France. (*nous*, +2)

3. Approches des cooccurrences de *JE*

3.1. *LEXICO3 : la cooccurrence par un chemin inattendu : les spécificités*

L'outil « carte des sections » établit la distribution de *je* dans la linéarité du texte, délimité ici en paragraphes, dressant une « topographie textuelle » du corpus. (Lamalle, Salem, 2002). Le calcul des cooccurrents spécifiques met en évidence à partir des sections ainsi délimitées les formes spécifiques des paragraphes attestant la forme. La liste produite porte sur l'ensemble du corpus et ne présente que les formes dont l'indice de spécificité est supérieur à 2. Le seuil de probabilité est de 5 %, la fréquence minimale des formes considérées est de 2 occurrences (Tableau 1).

Les spécificités positives montrent qu'une forte proportion de verbes, gravite autour du référent du locuteur : verbes marquant la volition (*souhaite, voudrais, forme* [le vœu], *veux),* le jugement (*pense, crois),* factifs (*fais* +3), verbes d'état et auxiliaires (*suis*, *ai*), énonciatifs (*dis*), verbes marquant la connaissance (*sais*, *vois*), quelques infinitifs (*Dire, adresser*), constituent l'essentiel du système verbal restitué par la recherche des cooccurrents spécifiques. On note par ailleurs la présence d'un verbe conjugué à la troisième personne du pluriel : *vivent*

(+3), dont on trouve la réalisation dans de fréquentes adresses aux Français *qui vivent à l'étranger* (De Gaulle, 1967), *qui vivent dans la solitude* (V.G.E, 1978)*, qui vivent dans la peine* (Mitterrand, 1986)*, qui vivent dans la difficulté quotidienne* (Mitterrand, 1988)...

Ces messages sont particulièrement marqués par des verbes de « circonstance » (*souhaiter*, *adresser former*), par des volitifs et des verbes exprimant la connaissance. Cependant, cette classification sémantique établie a priori ne résiste pas longtemps à l'examen des contextes.

Une analyse approfondie indique que la forme *voudrais* est intimement liée au référent de l'interlocuteur, sur employée chez les locuteurs qui précisément multiplient les marques énonciatives en direction des Français. La valeur n'est donc que rarement purement volitive, les emplois étant essentiellement métadiscursifs, intervenant dans des annonces de plan où bien souvent le locuteur s'adresse à une certaine catégorie de Français (*Je voudrais d'abord exprimer ma sympathie à toutes celles et à tous ceux qui vivent ces derniers jours de 1999 dans l'épreuve.* [Chirac, 1999]). Giscard et Pompidou qui entretiennent un lien très étroit avec les Français emploient cette forme dans une modalité directive qui intensifie la relation (*Je voudrais que vous sentiez, que vous compreniez…*).

Les contextes de *veux* montrent une tendance vers des emplois métadiscursifs ou explicatifs (*Je veux dire*), même si la volition apparaît parfois chez Chirac et Mitterrand dans une faible mesure.

Parmi les verbes exprimant le jugement, *penser* intervient essentiellement dans des énoncés énumératifs (*Je pense aux artisans, je pense aux agriculteurs, je pense à certaines petites entreprises*) mais bien souvent affectifs et empathiques, liés à l'évènementiel (*Et je pense aussi à nos compatriotes de Toulouse…*[Chirac, 2001]) ou plus généralement destinés à adresser un geste en direction des Français les plus démunis, évocation qui devient systématique à partir de Pompidou. (*Je pense spécialement à ceux de nos aînés qui vont franchir seuls le cap du nouvel an.* [Chirac, 2000], *Je pense à celles et à ceux d'entre vous qui connaissent le deuil, les chagrins, le poids de la maladie et de la solitude, qui souffrent du chômage*. [Mitterrand, 1981]). Parmi les verbes exprimant la conscience et la connaissance, l'examen des contextes montre que la forme *sais* entre essentiellement dans des modalités allocutives. Les emplois soulignent l'empathie, plus particulièrement chez Chirac, souvent comme procédé argumentatif. Cette marque d'empathie introduit dans de nombreux cas chez Chirac une relance incitative et mobilisatrice (Tableau 2).

Le champ lexical du rituel, verbes, substantifs, formules d'adresse est largement représenté parmi les formes qui gravitent autour du *je* présidentiel :*compatriotes*, *chers*, *mer* (*compatriotes d'outre mer*), *Français*, *Françaises*, *vive*, *année*, *bonne*, *heureuse*, *vœux*.

Parmi les spécificités négatives – minoritaires et dont les indices se révèlent moins élevés que pour les sur-emplois – nous retiendrons le diagnostic porté sur la première personne du pluriel. (*nous*, -7). Ainsi, le schéma *je/vous*, dont on relève les traces parmi les pronoms personnels et adjectifs possessifs se construit par le rejet de la première personne du pluriel.

3.2. *La notion d'environnement thématique HYPERBASE*

C'est à partir de la recherche des contextes qu'il nous est offert d'étudier l'environnement thématique sous Hyperbase. Cette recherche thématique procède d'un calcul de spécificité particulier. Plutôt que de porter un jugement sur la fréquence d'apparition d'une forme dans une sous partie du corpus par rapport aux autres et à l'ensemble, on calcule les spécificités des mots qui se trouvent au voisinage d'une forme pôle. En d'autres termes, et nous empruntons cette formulation à Etienne Brunet : « Ce programme de repérage thématique fait le décompte de tous les mots situés dans le même paragraphe que le ou les mots-pôles et mesure leur spécificité, c'est-à-dire la plus ou moins grand attirance que le mot-pôle exerce sur eux. » Le choix des partitions n'est pas sans incidence sur la recherche thématique. Aussi adopterons-

nous la partition locuteur qui présente l'avantage de ne pas morceler le corpus par trop ainsi qu'un découpage en paragraphes qui constitue un échantillon contextuel suffisamment large et correspond à une division naturelle du corpus.

Les formes constitutives de « l'environnement thématique » de *je* sont classées selon leur rang de significativité, c'est-à-dire dans l'ordre décroissant de l'écart réduit, qui s'étend ici de la valeur 25,79 à 2. Contrairement aux cooccurrents spécifiques de Lexico3, aucun diagnostic n'est porté sur les sous emplois. La seconde colonne présente l'effectif de la forme considérée sur l'ensemble du corpus, la dernière colonne indique le nombre d'occurrences dans la sous partie, constituée à partir des contextes de *je*. Pour faciliter la lecture, et la comparaison avec les résultats produits par Lexico3 nous avons porté les indices de spécificité relevés par cet outil. (Colonne *Sp*) (Tableau 3).

L'essentiel des faits que nous avions abordés par l'entrée des cooccurrents spécifiques est restitué. Les premiers rangs sont identiques à ceux observés au moyen de la méthode de cooccurrents spécifiques, ce qui conforte nos premières conclusions : le rituel et la thématique des vœux prédominent. Parmi les verbes : *souhaite* dont on a déjà examiné les contextes et établi qu'il était essentiellement employé dans la formulation de vœux, de même *forme*, *adresser*, mais aussi *limiter* (*je ne voudrais pas me limiter à vous présenter mes vœux…)*. Les usages de la première personne sont essentiellement conditionnés par le genre discursif y compris chez les locuteurs dont l'énonciation est peu personnalisée. Le *Je* est à la fois dialogique et familier, rituel et circonstanciel. On s'interrogera sur la procédure d'élagage appliquée sous Hyperbase – dont on maîtrise peu les paramètres – mais dont la présentation synthétique fait émerger les faits saillants du corpus mais aussi des données plus anecdotiques, qui apparaissaient sous Lexico parmi de nombreuses autres formes, dotées généralement d'indices très faibles de spécificité. On soulignera cependant l'absence remarquable des formes *chers* (indice positif de 4 sous Lexico) , *compatriotes* (+5), *France* (+3), *Vive* (+5) *bonne* (+5) qui entrent dans la réalisation de formules d'adresse ou qui appartiennent à la dimension rituelle du discours (*Vive la France, bonne année…*). Ces outils fournissent deux représentations des faits de cooccurrences, propres à orienter l'analyse dans des directions parfois différentes, d'où la nécessité de les interroger systématiquement.

3.3. *Du calcul hypergéométrique aux lexicogrammes (WEBLEX)*

Le calcul de cooccurrences implémenté dans Weblex repose sur le modèle développé par Pierre Lafon (1984). On se reportera à Heiden (2004) et Tournier (2003) pour des approfondissements méthodologiques.

« Le lexicogramme d'un mot s'interprète comme une synthèse des cooccurrents gauches et droits d'un mot, à l'intérieur de toutes les phrases où il apparaît » (Heiden, 2004). Le tableau 6 fournit ainsi une dimension supplémentaire qui ne pouvait être appréhendée avec les outils précédents et présente les principaux cooccurrents gauches et droits de la forme pivot *je* dans l'ensemble du corpus vœux au seuils de fréquence et co-fréquence minimale de 3 occurrences, probabilité de 5 % (soit 5.0^{e}-2), distance moyenne 1000 occurrences. Pour chaque cooccurrent la fréquence totale de la forme dans le corpus (**F**), sa co-fréquence avec la forme pôle, c'est-à-dire le nombre de rencontres attestées (**CF**), le diagnostic de probabilité de la rencontre (**P**) et la distance moyenne ($\mathbf{d_m}$) fournissent les indications quantitatives et statistiques de ces rencontres.

Nous constatons au premier regard un déséquilibre entre le cooccurents gauches et les cooccurents droits qui dominent très largement, ce qui ne saurait surprendre compte tenu de la nature de la forme pivot qui implique par essence de plus fréquentes relations sur sa droite que sur sa gauche. Les cooccurrents gauches restituent la dimension rituelle du discours. Cinq d'entre eux sont constitutifs de formules d'adresse (*mes chers compatriotes de métropole et d'outre mer*. La position du substantif « vœux » ne surprendra pas d'avantage : (*les vœux que*

je forme, *que je vous adresse*…) ni la présence de la forme *cœur* dont les contextes attestent la réalisation (*C'est de tout cœur que je*...).

Les verbes, généralement post-posés au pronom personnel figurent tout naturellement à droite du pivot. *Souhaite* se trouve au premier rang (tri par probabilités) comme c'était le cas avec les autres analyses. Les onze premières formes à droite sont des verbes dont la morphologie indique qu'ils sont fléchis à la première personne du singulier, et qui sont peu distants de la forme pivot. On notera que le temps est majoritairement le présent de l'indicatif. Après la catégorie verbale viennent des adjectifs et des substantifs qui ancrent le discours dans le présent, ou sont issus de la thématique des vœux : (*heureuse*, *vœux*, *soir*, *bonne*, *année*, *adresse*)…

Peu de termes politiques émergent de ces lexicogrammes, ainsi que nous l'avions déjà observé au moyen des autres outils cooccurentiels : le *je* est essentiellement mobilisé par le genre discursif. Une dernière remarque concerne la distance moyenne des cooccurrents, plus importante sur la gauche de la forme pivot que sur sa droite où l'on observe quelques collocations. Qu'en est il de la présence forte de l'interlocuteur que nous avions cru déceler dans les analyses précédentes ? Parmi les verbes, aucun ne semble être conjugué à la deuxième personne du pluriel, ainsi que nous l'avions déjà noté. On ne s'étonnera pas de l'absence des marques personnelles renvoyant à l'interlocuteur dont on a noté la forte spécificité sous Lexico, confirmée par Hyperbase. Nous avons en effet choisi de conserver l'élagage des formes outils du vocabulaire afin de porter un regard différent sur le corpus, en ne considérant que les seuls mots pleins. Une expérience menée sur ces mêmes lexicogrammes sans suppression de ces formes a cependant confirmé l'attirance importante des marques de la première du singulier et de la seconde du pluriel.

Le lexicogramme récursif (Tableau 4) associé à une forme révèle des réseaux de cooccurrences sur le principe d'un enchaînement de cooccurrents de cooccurents. Partant de la source du lexicogramme (*je*) il en établit les principales attirances. À partir de ces nouvelles formes, d'autres cooccurents sont mis en évidence selon le principe que chaque nouveau cooccurrent est à son tour pris comme source. La représentation qui en résulte synthétise l'ensemble des connexions qui prennent naissance autour de la forme pôle. Le lexicogramme récursif de la première personne du singulier, a été établi avec élagage des formes outils aux seuils précédemment adoptés.

La thématique des vœux ressort nettement de cette représentation. La connexion de verbes gravitant autour du *je* est très visible : *forme*, *adresser*, *sais*, *demande*, *dis*…, verbes qui nous sont désormais familiers. Dans cette configuration, l'ordre des termes n'est pas restitué bien qu'il soit possible de le faire. Cette perspective spatiale permet de saisir des faits qui ne seraient pas nécessairement apparus aussi clairement sur des listes hiérarchisées. Ainsi les formules d'adresse sont-elles matérialisées par les nœuds *départements*, *territoires*, *outre-mer*. Les qualifiants de *vœux*, *ardents* et *confiants* que le général de Gaulle *adresse* à la France sont clairement identifiés, de même que les emplois métadiscursifs (*je veux dire*, *je voudrais dire*, *je voudrais adresser*). On se gardera cependant d'interpréter trop hâtivement ces connexions. Suivons par exemple la « branche » qui relie le pronom personnel à *disais*. Le verbe est lui-même relié à *an* qui forme un nœud à partir duquel deux nouveaux arcs partent vers *fêter* et *nouvel*. Nul doute que le substantif *an* entre en cooccurrence avec le verbe et l'adjectif. Des réalisations comme « *fêter le nouvel an* » ne seront pas difficiles à attester. Pas de doute non plus sur la relation *je-disais* vérifiée en contexte. Il en va tout autrement de réalisations poly-cooccurrentes qui attesteraient *je*, *disais*, *fêter*, *nouvel an* dans un même contexte. Ceci tient au caractère de récursivité des lexicogrammes selon lequel se révèlent les principales formes cooccurrentes de la marque personnelle – dont *disais* – prises à leur tour comme sources – révélant le substantif *an* – jusqu'à saturation du lexicogramme – aboutissant à *fêter* et *nouvel*. Chaque nœud matérialise ainsi des paliers de cooccurrences dont les réalisations

n'entrent pas nécessairement dans les mêmes contextes. Il s'agit donc, bien évidemment, de ne pas considérer le chemin « artefactuel » que nous venons de retracer comme un schéma syntaxique, un « squelette de phrase » (Martinez, 2003).

3.4. *Une utilisation particulière du tri croisé (ALCESTE)*

L'expérimentation menée ici repose sur une utilisation un peu particulière d'Alceste. Il ne s'agit plus de faire émerger les structures saillantes du corpus, d'en identifier les classes d'énoncés ou « mondes lexicaux » (Reinert, 1993 et 1998) mais d'utiliser le tri croisé pour mettre en évidence les cooccurrents du *je*, sur l'ensemble du corpus.

Cette analyse produit une classification binaire selon la présence absence de la forme croisée avec le corpus au sein des énoncés. La totalité des unités de contexte élémentaire est prise en compte sur la base d'une classification ascendante, 100 % des U.C.E sont donc classées, contrairement aux analyses usuelles qui reposent sur le principe de la classification descendante.

Le tableau 6 présente les formes les plus caractéristiques des deux classes obtenues, ordonnées selon le Khi2 décroissant. Ces classes nous donnent confirmation de certains faits observés par ailleurs.

– Sur l'aspect pronominal, la dimension dialogique du discours est confortée, ainsi que le rejet de la première personne du pluriel: *vous* figure au second rang des formes significatives de la classe 1, les marques de la première du pluriel à un rang identique dans la classe opposée.

– Sur l'aspect énonciatif: Parmi les énoncés qui s'opposent au *Je* la variable loc_dg est la première forme significative de la deuxième classe. C'est-à-dire que la plupart des énoncés du locuteur de Gaulle se trouve classé dans la catégorie la plus éloignée du *je*. De Gaulle étant le locuteur qui use le moins de cette personne, ce résultat tend à conforter nos premières conclusions.

– Sur la chronologie: D'autres variables figurent parmi les premières classes constitutives de la classe 2. Les années 63, 65, 61, 68, 60, 66, 2001 et 2000 sont en effet dotées d'indices de Khi2 relativement importants. Ainsi cette classification binaire et somme toute assez brutale porte un éclairage chronologique sur les emplois pronominaux puisque, recoupant cette information par les spécificités par partie, il ressort que la forme *je* est sous employée dans les messages correspondants par rapport à l'ensemble du corpus. Ainsi résumons-nous notre démarche par ce va et vient entre les différents outils, les résultats de l'un nous engageant à interroger l'autre.

Dans la première classe, en revanche, la variable *Giscard* est la plus représentative. Nous avions constaté qu'il était celui chez qui la personnalisation du discours était la plus sensible. Ses messages par ailleurs sont ceux qui laissent la part la plus importante à la thématique des vœux. Ce fait se vérifie, là encore par la méthode des spécificités.

Cette thématique des vœux est précisément très représentée dans la première classe, à travers des substantifs : *vœux*, *année*, *soirée* à travers des adjectifs : *heureux*, *bonne* mais aussi des verbes : *souhait*<, *forme*, *adresser*, *présenter*, *échanger*. Le rituel et les formules d'adresse sont par ailleurs constitutifs de cette première classe : *Compatriote* +, *cher* +, *metropole* +,. On notera que le lexique est plus affectif dans cette même classe, tandis que la classe 2 recense un lexique plutôt politique, économique et social: *economie*, *développement*, *organisation*, *conflit*…

Cela nous amène à nous interroger sur la nature des relations entretenues entre les marques de la première personne et les différentes thématiques des messages. : Le *je* rejette-t-il dans ces

messages la dimension politique au profit d'un lexique plus circonstanciel ou rituel. Cette dimension politique serait-elle alors développée autour d'autres référents ?

4. Des résultats convergents, mais des mises en reliefs contrastées

Le croisement des approches et des outils statistiques dont la finalité première n'est pas nécessairement la recherche localisée de phénomènes cooccurentiels révèle des résultats convergents pour les faits les plus saillants. Cette démarche expérimentaliste et comparative a permis de réunir un faisceau sur lequel l'analyse interprétative peut s'appuyer plus fortement, que sur les faits révélés par un outil unique. Le croisement des perspectives valide par ailleurs la démarche lexicométrique sur un corpus de faible étendue tel que celui des vœux présidentiels. L'application de paramètres et de seuils, plus ou moins intuitifs mais souvent différents, le choix des fenêtres contextuelles, l'élagage ou non de certaines formes de vocabulaire, sont autant de facteurs qui incitent à confronter ces outils avant toute phase d'interprétation. Spécificités positives et négatives, topographie textuelle (Lexico 3), présentation synthétique (Hyperbase), ou ordonnée (Weblex lexicogrammes simples), réseaux d'affinités et récursivité (Lexicogrammes récursifs), partition binaire et classification ascendante (Alceste), fournissent des perceptions différentes et complémentaires des faits de cooccurrences.

Références

Chateauraynaud F. (2003). *Prospéro – Une technologie littéraire pour les sciences humaines*. CNRS Editions.

Heiden S. (2004). Interface hypertextuelle à un espace de cooccurrences: implémentation dans Weblex. In *Actes des JADT 2004*.

Labbé D. (1990). *Le vocabulaire de François Mitterrand*. Presses de la Fondation Nationale des Sciences Politiques.

Labbé D. (2003). *Corneille dans l'ombre de Molière. Histoire d'une découverte*. Les Impressions Nouvelles, Collection « Bâtons rompus ».

Lafon P. (1984). *Dépouillements et Statistiques en Lexicométrie*. Slatkine-Champion.

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des JADT 2002.*

Martinez W. (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*. Thèse de troisième cycle.

Reinert M. (1993). Les « mondes lexicaux » et leur logique. *Langage et société*, vol. (66).

Reinert M. (1998). Quelques interrogations à propos de l'« objet » d'une analyse de discours de type statistique et de la réponse « Alceste ». *Langage et société*.

Reinert M. (1990). Système Alceste: Une méthodologie d'analyse des données textuelles. In *Actes des JADT 1990.*

Tournier M. (2003). *De France à Je. La traversée des emplois Cooccurrences et connexions*. Des sources du sens, École Normale Supérieure Lettres Sciences Humaines Lyon, Collection Langages.

Annexes

Forme	Frq. Tot.	Fréquence	Coeff.	Forme	Frq. Tot.	Fréquence	Coeff.
je	344	344	51	j	88	51	3
souhaite	65	64	23	grandeur	7	7	3
vous	326	227	23	soir	42	28	3
voeux	80	62	10	adresser	7	7	3
mes	102	75	10	ma	20	15	3
sais	19	19	8	nom	30	21	3
voudrais	19	19	8	mon	29	19	3
pense	22	21	7	amis	11	10	3
suis	26	24	7	fraternité	19	14	3
forme	17	16	5	seuls	14	12	3
vive	60	42	5	france	302	150	3
heureuse	22	19	5	vivent	10	9	3
bonne	76	51	5	m	23	17	3
veux	11	11	5	vois	6	6	3
crois	11	11	5	ministre	6	6	3
que	677	336	5	famille	25	17	3
dire	48	35	5	très	27	18	3
dis	12	12	5	fais	6	6	3
compatriotes	62	43	5	françaises	41	27	3
ai	41	31	5	doivent	14	1	-3
année	205	110	4	quel	12	0	-3
vos	39	27	4	la	1397	546	-3
chers	55	37	4	algérie	21	3	-3
me	22	17	4	économique	46	10	-3
français	142	80	4	qu	313	108	-3
votre	59	38	4	peut	50	11	-3
chacune	25	20	4	europe	99	28	-3
mer	19	14	3	part	32	3	-5
				nous	655	217	-7

Tableau 1. Cooccurrents spécifiques de « je » sur la totalité du corpus « vœux »

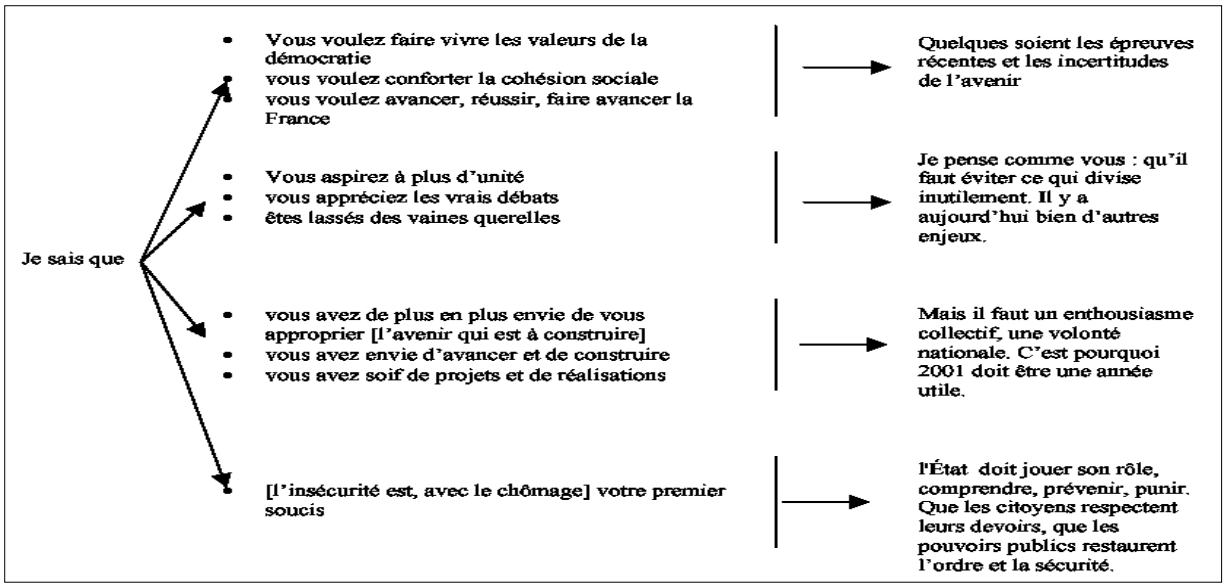


Tableau 2. Marques de l'empathie et relances incitatives chez J. Chirac

Ecart	Corpus	Extrait	Mot	Sp.	Ecart	Corpus	Extrait	Mot	Sp.
25.79	344	347	JE	51	2.74	4	4	LÉGISLATIVES	2
11.80	323	213	VOUS	23	2.74	4	4	AUGMENTÉ	2
11.06	65	65	SOUHAITE	23	2.71	29	17	MON	3
7.23	79	58	VŒUX	10	2.65	11	8	RAISONS	2
6.46	19	20	VOUDRAIS	8	2.61	19	12	FRATERNITÉ	3
6.17	26	24	SUIS	7	2.60	127	58	AUX	1
5.99	22	21	PENSE	7	2.60	122	56	AVEC	2
5.98	19	19	SAIS	8	2.60	17	11	ADRESSE	2
5.56	48	35	DIRE	5	2.50	6	5	UNS	2
5.50	41	31	AI	5	2.50	6	5	MINISTRE	3
5.44	674	301	QUE	5	2.50	6	5	EXPRIME	1
5.15	17	16	FORME	5	2.50	6	5	DÉPARTEMENTS	2
4.75	12	12	DIS	5	2.50	6	5	AVAIS	2
4.55	11	11	VEUX	5	2.40	178	77	PAS	2
4.55	11	11	CROIS	5	2.40	33	18	GOUVERNEMENT	1
4.34	25	19	CHACUNE	4	2.40	8	6	COMPRENDRE	2
4.25	59	36	VOTRE	4	2.35	202	86	ANNÉE	4
4.20	22	17	ME	4	2.35	18	11	AVEZ	2
4.17	101	55	MES	10	2.33	14	9	SEULS	3
4.03	42	27	SOIR	3	2.33	14	9	CHOSES	2
3.68	30	20	NOM	3	2.22	41	21	FRANÇAISES	3
3.63	7	7	ADRESSER	3	2.17	90	41	AUSSI	2
3.59	3	4	LIMITER	2	2.16	21	12	POSSIBLE	1
3.55	101	52	CEUX	2	2.14	30	16	ABORD	2
3.52	39	24	VOS	4	2.13	37	19	AN	2
3.49	27	18	CŒUR	2	2.13	5	4	VRAIMENT	1
3.36	6	6	VOIS	3	2.13	5	4	VENU	1
3.36	6	6	FAIS	3	2.13	5	4	STRASBOURG	1
3.30	22	15	BONHEUR	2	2.13	5	4	RÉPONDRE	1
3.28	11	9	AMIS	3	2.13	5	4	PROGRESSER	1
3.08	25	16	FAMILLE	3	2.13	5	4	OFFRE	1
3.08	23	15	M'	3	2.13	5	4	KOWEÏT	2
3.07	5	5	MESSAGE	2	2.13	5	4	GARANT	1
3.07	5	5	MARS	2	2.13	5	4	FRAPPÉ	1
3.07	5	5	FIER	2	2.13	5	4	FIDÈLE	1
3.07	5	5	FAIBLES	1	2.13	5	4	ADRESSENT	1
3.07	5	5	DISAIS	2	2.12	19	11	MER	3
3.05	83	42	J'	3	2.09	17	10	ÉTRANGER	2
3.01	10	8	VIVENT	3	2.04	7	5	PRÉSIDENT	1
2.88	26	16	TRÈS	3	2.04	7	5	PARLER	1
2.85	22	14	HEUREUSE	5	2.04	7	5	MAJORITÉ	2
2.85	20	13	MA	3	2.04	7	5	FRATERNELLE	1
2.84	7	6	GRANDEUR	3	2.02	11	7	SOUFFRENT	1
2.84	7	6	DEMANDE	2	2.02	11	7	PROFESSIONNEL	2
2.77	142	65	FRANÇAIS	4	2.02	9	6	SOLITUDE	1
2.74	4	4	VAIS	2	2.02	9	6	RÉUSSIR	1
2.74	4	4	TIRE	2	2.02	9	6	DÉPIT	1
2.74	4	4	TIENS	2	2.02	9	6	DÉBUT	2
2.74	4	4	REÇU	2	2.00	24	13	TROP	2
2.74	4	4	QUICONQUE	2	2.00	24	13	BESOIN	1

Tableau 3. Environnement thématique de la forme « je ». Partition locuteur,contexte paragraphe

je

(346)**cooccurrents gauches**					**cooccurrents droits**									
	f	**cf**	**p**	**d_m**		**f**	**cf**	**p**	**d_m**					
chers	55	21	4e-06	7.3	souhaite	65	64	6e-57	0.2	faibles	5	4	2e-03	20.8
métropole	18	10	3e-05	11.1	pense	22	21	9e-18	0.0	salue	3	3	3e-03	0.7
compatriotes	62	20	1e-04	7.2	voudrais	19	19	3e-17	0.1	dirai	3	3	3e-03	1.0
françaises	41	15	2e-04	10.7	sais	19	18	4e-15	0.6	répète	3	3	3e-03	1.3
outre-mer	14	7	1e-03	8.6	forme	19	16	8e-12	0.0	souviens	3	3	3e-03	1.3
voeux	81	21	2e-03	4.9	crois	11	11	3e-10	0.1	assure	3	3	3e-03	14.0
coeur	27	9	8e-03	4.3	veux	11	11	3e-10	0.1	sûr	12	6	3e-03	1.0
vois	6	3	4e-02	17.0	dis	12	11	3e-09	0.7	rendre	9	5	4e-03	7.6
					dire	39	19	1e-07	3.3	promis	6	4	4e-03	2.0
					adresser	7	7	9e-07	6.9	espère	6	4	4e-03	5.8
					fais	6	6	7e-06	0.8	propose	6	4	4e-03	6.5
					heureuse	22	12	7e-06	13.5	dit	14	6	8e-03	3.8
					voeux	81	26	9e-06	7.3	fraternelle	7	4	9e-03	17.8
					soir	42	17	1e-05	6.8	demandé	4	3	9e-03	2.0
					bonne	75	24	2e-05	8.5	satisfaction	4	3	9e-03	12.0
					demande	7	6	4e-05	2.2	tire	4	3	9e-03	19.3
					disais	5	5	5e-05	0.8	nom	30	9	2e-02	6.4
					vois	6	5	3e-04	0.0	vivre	30	9	2e-02	13.1
					vais	4	4	4e-04	0.0	parle	8	4	2e-02	2.2
					année	202	42	7e-04	10.1	constate	5	3	2e-02	0.0
					adresse	17	8	9e-04	1.0	ardents	5	3	2e-02	3.0
					fier	5	4	2e-03	6.5					

Tableau 4. Lexicogramme de la forme je dans le corpus vœux. Seuils : f 3, cf 3, p 5.0E-2, d_m 1000.0

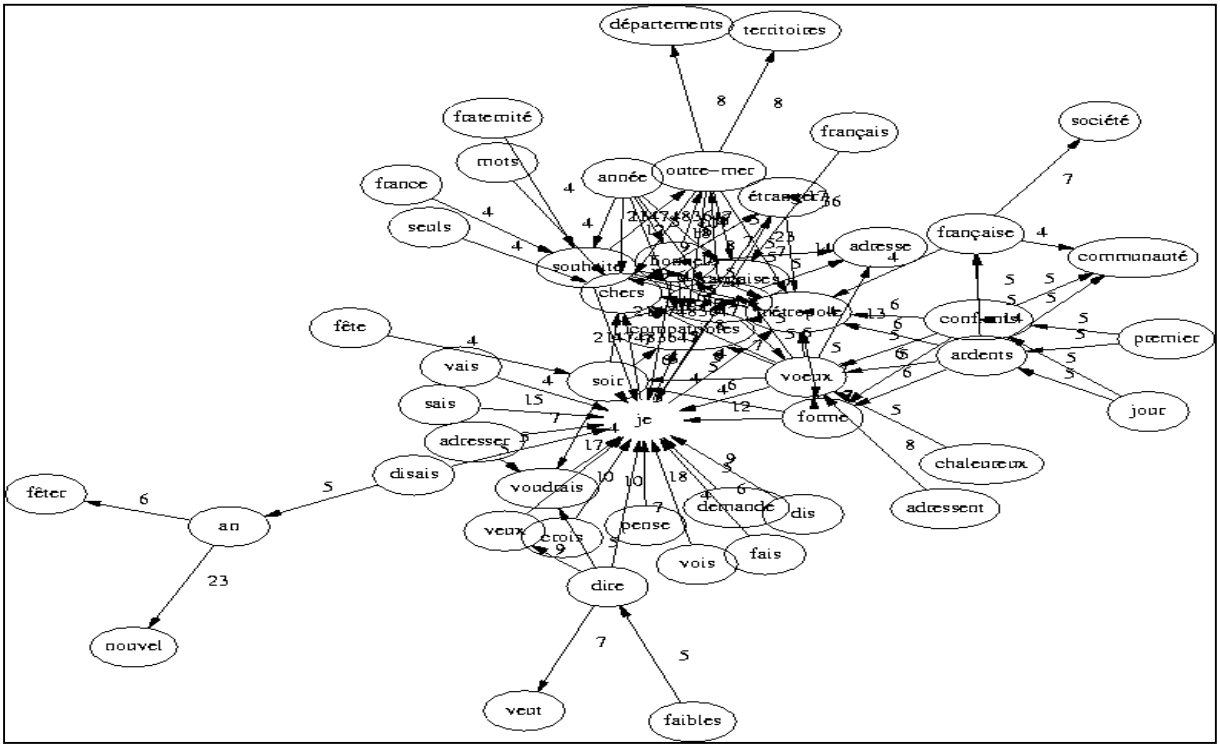


Tableau 5. Lexicogramme récursif de la forme je dans le corpus vœux.
Seuils : **f** 3, **cf** 3, **p** 5.0E-2, $\mathbf{d_m}$ 1000.0

Classe 1: 255 U.C.E (27%)

Forme réduite	Khi2
je	969
vous	148,97
souhait<	148,06
suis	60
dire.	45,77
mes	44,38
voeu+	38,95
dire+	38,14
forme+	37,03
me	27,54
annee+	27,44
vouloir.	27,02
heur+eux	23,87
compatriote+	23,61
francais+	23,31
adress+er	22,45
cher+	19,82
ai	18,98
nom+	17,59
outre	15,11
qu+	14,12
enjeu+	14,07
adresse+	13,15
*loc_giscard	12,76
propos+er	12,36

Forme réduite	Khi2
votre	12,07
soir+	12,07
bonne+	11,34
redire.	11,25
tiens	11,25
present+er	11,25
echang+er	11,25
metropole+	10,98
savoir.	10,9
m	10,52
soiree+	10,12
avais	10,12
coeur+	9,97
promettre.	9,86
parl+er	9,43
année_1974	9,34
mon	9,34
pas	9,32
ce	9,27
ma	8,67
mer+	8,67
recu+	8,43
aim+er	8,43
demand+er	8,09
rappel+er	7,63

Classe 2 : 714 U.C.E (73%)

Forme réduite	Khi2
*loc_dg	52,38
nous	14,84
econom+	10,6
année_1963	10,31
devoir.	8,67
année_1965	7,29
année_1961	7,27
année_1968	6,24
année_1960	6,16
moyen+	5,56
developpement+	5,5
*année_1966	5,12
arme+	5,07
cooperati+f	5,07
techn+	5,07
peuple+	5,06
face+	4,71
organisat+ion	4,71
*année_2001	4,57
avons	4,52
elle	4,46
jusqu+	4,34
plan+	4,34
conflit+	4,34
inflation<	4,34

Forme réduite	Khi2
trouv+er	4,1
ailleurs	3,97
leur	3,97
guerre+	3,93
moderne+	3,73
solid+e	3,61
rapport+	3,61
acquerir	3,61
soi	3,61
quel	3,61
troubl+er	3,61
scientifi<	3,61
ensemble+	3,55
*année_2000	3,55
*année_1962	3,44
mondia+	3,44
moment+	3,38
cas	3,24
dehors	3,24
telle	3,24
multipli+er	3,24
monetaire+	3,24
but+	3,24
route+	3,24
aid+er	3,01

Tableau 6. Croisement de la variable « je » avec l'ensemble du corpus (Alceste)

Représentations des réseaux de mots associés

Christophe Lejeune

FNRS – Sociologie Générale – Université de Liège
Bd du Rectorat 7 B32 Bte 47 – Belgique
christophe.lejeune@ulg.ac.be

Abstract

The paper presents networks of associated words as it is used in qualitative data analysis software such as Candide, Réseau-Lu T-Lab on one side and Prospéro on the other side. The precise algorithm of both Candide and Prospéro is described.
Strength and limits of such a feature are compared with the gain it provides to social scientists.
The paper concludes on a procedure helping to judge the adequacy and accuracy of categories of analysis constructed by the researcher.

Résumé

Cet exposé présente les réseaux de mots associés utilisés dans les logiciels d'analyse des données qualitatives comme Candide, Réseau-Lu et T-Lab ou comme Prospéro. L'algorithme de Candide et Prospéro est précisément décrit.
Les contributions et les limites de ce type de fonctionnalités sont jaugées en fonction de les apports qu'elles fournissent aux chercheurs en sciences sociales.
L'exposé se termine sur une procédure permettant d'éprouver la pertinence et la validité des catégories d'analyse construites par le chercheur.

Mots-clés : Candide, Prospéro, CAQDAS, cooccurrence, cluster, carte de lien.

1. Délimitation du champ

Mon propos porte sur les outils logiciels qu'utilisent les sciences sociales lorsqu'elles s'adonnent à l'analyse qualitative[1]. Sont exclus de ce champ les outils quantitatifs provenant de l'analyse factorielle ou des statistiques (SAS, Spss, R). Restent donc les outils qualitatifs dont les plus connus sont employés par les chercheurs du monde entier et désignés par l'acronyme CAQDAS (pour Computer-Assisted Qualitative Data Analysis Software). Parmi ces outils figurent en bonne place Nud*Ist, AtlasTi, Kwalitan, The Ethnograph, Intext (à ne pas confondre avec Intex) et bien d'autres[2]. Je situe l'origine des outils d'analyse qualitative informatisée dans les années soixante avec The General Inquirer ; un logiciel initialement développé chez IBM, notamment par Philip Stone, et toujours en activité aujourd'hui.

S'agissant d'outiller l'analyse compréhensive de données textuelles, j'ai également restreint mon intérêt aux outils dédiés à la langue française. Cela écarte les outils cités ci-dessus. La

[1] Je remercie pour leur aide Jean-Pierre Charriau et Francis Chateauraynaud (Doxa) ainsi que le Fonds National de la Recherche Scientifique (FNRS) qui a rendu cette étude possible.

[2] Je renvoie le lecteur intéressé à la rubrique de l'Open Directory Project dédiée à ces outils que je tiens personnellement à jour : http://dmoz.org/Science/Social_Sciences/Methodology/Qualitative/Tools/ .

raison réside dans les différences linguistiques et culturelles transportées par des énoncés provenant d'Angleterre ou de Belgique.

Le champ des logiciels concernés compte donc des outils comme Alceste, Candide, Civilité, Hyperbase, Leximappe, Prospéro, Réseau-Lu, Sampler, Sato, Sphinx, T-Lab et Tropes[3]. Dans ces logiciels, plusieurs techniques sont présentes qui permettent de rendre compte des corpus analysés (Jenny, 1997). Je me propose de me pencher sur les réseaux de mots associés (RMA). Ce type de construction est très apprécié car il se double souvent d'une représentation en deux dimensions sous forme de graphes. À la différence d'une arborescence, ce graphe réticulaire est non dirigé, il provient de la génération d'agrégats de mots fortement liés (la métrique dépend de la récurrence de leur cooccurrence). Ces agrégats (ou clusters) permettent de dégager des pôles (représentés par les nœuds) correspondant aux associations les plus fréquentes, donc, selon les utilisateurs de ces cartes (Callon *et al.*, 1991b ; Courtial et Kerneur, 1995 ; Teil, 1991 ; Teil, 1994), aux différents grands thèmes du corpus ; toujours selon le même principe de calcul, ceux-ci apparaissent à leur tour liés (ou non) entre eux.

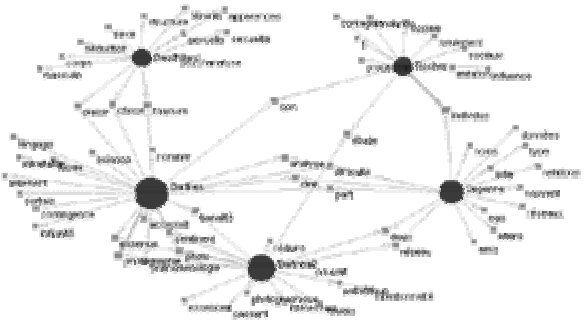


Fig. 1. Réseau-Lu

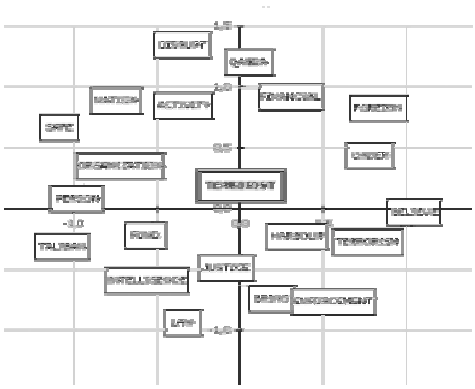


Fig. 2. T-lab

Les différents algorithmes qui permettent de construire ce type de réseau se basent sur la cooccurrence. Ce faisant, c'est le contenu des textes (et même plus précisément celui des énoncés) qui est analysé. J'écarte donc les réseaux de co-citations (Callon, 1993 ; Flake *et al.*, 2002) ou les réseaux de mots-clés pour conserver uniquement les techniques calculant un indice de cooccurrence déduit de données internes aux textes. Je rappelle que les techniques qui m'intéressent s'insèrent dans une démarche scientifique ; dès lors, l'exploitation de réseaux de mots associés dans des applications de représentation non-normée de l'information ne seront pas abordées frontalement, même si leur algorithme ne diffère pas radicalement de celui utilisé par les sociologues. Cette restriction écarte des outils comme ceux que mobilisent les outils de recherche d'information sur Internet, comme l'ancienne interface du moteur Altavista, appelée LiveTopics[4] (Bourdoncle, 1997), celle du méta-moteur Kartoo ou celle du nouveau moteur australien Mooter. Contrairement aux outils scientifiques, ces outils de recherche d'information produisent des graphes non-normés : leurs arcs sont dépourvus de métrique (ce qui signifie que leur longueur n'a pas de signification).

[3] Pour une description de ces logiciels, voir mon site Internet : http://www.smess.egss.ulg.ac.be/lejeune/logiciels/

[4] Après l'abandon par Altavista de la technologie Refine/Cow9/Live, son concepteur a lancé Exalead qui double la génération automatique de mots-clés d'une base de données conçue « à la main ».

Enfin, les réseaux que je vais évoquer ne doivent pas être confondus avec les ontologies (que l'on désigne parfois par le terme de réseau conceptuel) ou encore les réseaux sociaux (qui illustrent des liens entre personnes et ne sont que très rarement construits à partir de textes, il s'agit d'ailleurs le plus souvent d'une technique relevant de l'analyse quantitative des données qui, en en tant que telle, est extérieure à mon propos).

2. Des justifications sociologiques, philosophiques et linguistiques

Divers paradigmes sociologiques, philosophiques et linguistiques recourent à des méthodes prenant une forme réticulaire. Les justifications de l'usage des réseaux peuvent donc se rapporter à ces trois registres. La théorie de la traduction offre une justification **sociologique** à cette technologie : les acteurs visant à intéresser leurs interlocuteurs (comme les promoteurs industriels, les inventeurs ou les scientifiques) traduisent les objectifs des autres comme congruents avec les leurs (Callon *et al.*, 1991b ; Latour, 1995 ; Callon, 1993). Ce faisant, ils tentent de se constituer en point de passage obligé (Callon, 1986). Ces intéressements se concrétiseraient dans l'assemblage des thèmes (ceux des autres et les leurs propres) dans leurs productions textuelles. Par conséquent, l'analyse des réseaux de mots associés conviendrait parfaitement pour suivre les traductions des acteurs.

Au niveau **philosophique**, la référence se tourne vers le courant empiriste du XVIIIe siècle avec Étienne de Condillac et plus encore David Hume dont l'analyse de l'idée même de causalité comme résultant de l'habitude d'une conjonction d'événements (sans que la consécution n'existe ailleurs que dans l'esprit du sujet) sert de justification aux adeptes de ces méthodes[5].

Enfin, le lien avec les **sciences de l'énonciation** réside dans la possibilité de déduire le sens du contexte, donc de rendre compte des phénomènes d'indexicalité. Les exemples les plus courants de cette détermination du sens par le contexte sont les anaphores et les déictiques, mais la pragmatique et l'ethnométhodologie ont montré que ce phénomène était bien plus large, au point que l'on puisse parler d'indexicalité généralisée ou absolue (Garfinkel et Sacks, 1970 ; Conein, 1993 ; Ducrot et Schaeffer, 1995 ; Ramognino, 1999). En outre, l'analyse d'un corpus de textes recourrant aux réseaux d'associations prend en considération le cotexte proche (celui de l'énoncé) ; cette proximité n'est cependant pas assimilable à l'entourage direct puisqu'elle tient compte des termes très éloignés au niveau syntagmatique mais néanmoins présents dans la même phrase, donc cooccurrents[6]. Par conséquent, la cooccurrence ne doit donc pas se confondre avec la distribution.

3. La génération des réseaux de mots associés

La réflexion critique sur la cooccurrence est le cœur de cette contribution. Il ne s'agit aucunement d'une nouveauté : elle est déjà présente à la fin des années cinquante (Osgood, 1959). J'évoque pour ma part l'utilisation de cette procédure au sein de deux types technologies littéraires : les cartes de liens (traitées dans ce troisième point) et Prospéro (qui est abordé dans le quatrième et dernier point).

Leximappe, Candide, Réseau-Lu et T-Lab mobilisent le calcul des réseaux de mots associés.

[5] Cette lecture de Hume est celle de Latour et Teil (1995).

[6] Je montre dans la troisième partie comment Prospéro introduit une « pondération de proximité » dans son calcul de la force d'association de deux termes.

Le simple comptage du nombre de fois que se présente la présence simultanée de deux termes (ou co-présence ou cooccurrence) donne une évaluation de la force du lien entre eux. Toutefois, ce type de comptage dépend trop de la fréquence. On utilise donc un indice statistique normédéfini par Bertrand Michelet et appelé *indice d'équivalence* (noté E_{ij}).

On calcule cet indice de la manière suivante :

$$E_{ij} = \frac{C_{ij}^{2}}{C_i \bullet C_j}$$

E_{ij} *désigne l'indice d'équivalence des mots i et j.*
C_{ij} *désigne la cooccurrence des mots i et j.*
C_i *désigne l'occurrence du mot i. Ce nombre est nécessairement supérieur ou égal à* C_{ij}.

L'indice d'équivalence évalue la force du lien : lorsqu'il vaut 0, les mots ne sont jamais présents ensemble ; s'il vaut 1, ils apparaissent toujours ensemble, on dit qu'ils sont équivalents (c'est, par exemple, le cas des mots composés). Les liens les plus forts sont réunis en agrégats (Callon *et al.*, 1991a).

Développé à partir de Leximappe™ par Geneviève Teil dans l'environnement Hypercard™, Candide™ ne tient pas compte des mots vides (articles, conjonctions, négations) et opère une lemmatisation des mots pleins. Afin d'éviter les ambiguïtés anaphoriques, les pronoms sont remplacés par leur antécédent (Ramaux, 1993).

Chaque unité de contexte élémentaire (qui correspond, dans le cas de ce logiciel, à la phrase) est transformé en réseau élémentaire grâce aux principes de calculs définis par Michelet. La somme de tous ces réseaux donne une représentation du texte.

À chaque mot est donc lié une série d'autres mots. La liste de ceux-ci s'appelle profil d'association (PA). La longueur du PA, c'est-à-dire le nombre de ses composantes, informe l'utilisateur sur le nombre de mots auxquels un terme est connecté. La comparaison des profils d'association est possible grâce à l'*indice de similitude* (noté S_{ij}) de Bertrand Michelet.

$$S_{ij} = \frac{PA_i * PA_j}{\|PA_i\| \bullet \|PA_j\|}$$

*« * » est le signe du produit scalaire; « • » est le signe de la multiplication.*
Le quotient désigne le produit scalaire normé des deux vecteurs.

Lorsque l'indice de similitude vaut 0, donc que les PA des deux mots considérés sont totalement hétérogènes, les deux mots appartiennent à des contextes différents.

Si l'indice de similitude vaut 1, les deux mots sont tous deux reliés aux mêmes mots. Si l'on considère que le sens d'un mot dépend donc de son contexte, on peut dire que les deux mots sont identiques au niveau de leur PA ; selon cette conception (le sens est contextuel), on peut considérer que le PA d'un mot en donne la définition.

Lorsque S_{ij} vaut 1, on se reporte à l'indice d'équivalence. S'il vaut également 1, les mots sont toujours associés *et* ils apparaissent toujours dans le même contexte. On peut donc les considérer comme un mot composé et les remplacer par un seul mot. C'est également la configuration des termes qui forment ce que l'on pourrait appeler, à la suite de Pierre Bourdieu, un couple épistémologique, et sont, à ce titre, tout à la fois antinomiques et complices, indissociables (Bourdieu *et al.*, 1968).

Par contre si E_{ij} vaut 0 (*et* que S_{ij} vaut 1), on a affaire à des mots toujours associés au même terme, qui ont « la même définition, mais n'apparaissent jamais ensemble. C'est le cas des synonymes » (Ramaux, 1993).

Candide fixe la frontière entre les agrégats lorsque se manifeste une « différence trop forte – supérieure à un seuil donné – entre la force de la dernière association prise en compte dans l'agrégat et la plus forte association suivante » (Teil, 1994).

Le degré de cooccurrence de deux concepts est proportionnel à leur proximité, tant thématique que graphique. Donc, deux mots fortement associés, dont l'indice d'équivalence (E) est élevé, sont proches ; l'arc qui les relie doit donc être court. La longueur des arcs (ou tenseurs) est par conséquent inversement proportionnelle à la valeur de E (Teil, 1991 ; Noyer, 1995b ; Jenny, 1997).

Cet outil permet de résoudre des cas de polysémie : un mot polysémique présente plusieurs « sous-profils d'association »; il est fortement lié à différents agrégats pas – ou peu – liés entre eux. On peut donc le scinder en deux mots (par exemple « avocat juriste » et « avocat fruit »). On peut utiliser le même raisonnement pour identifier les mots outils (ou mots vides) : il s'agit de mots fortement associés à l'ensemble des agrégats.

La lemmatisation et la suppression des mots vides (comme les négations), en tant que réduction sémantique et syntaxique, font que la simple présence d'un terme prime sur ce qu'il en est dit (Osgood, 1959 ; Jenny, 1997).

4. Prospéro

Le Programme de Sociologie Pragmatique et Réflexive sur Ordinateur[7] désigne une famille regroupant Ariel, Caliban, Chéloné, Marlowe, Sycorax et Tirésias (Chateauraynaud, 2003). Suivant la terminologie de Prospéro, les formes nominales et verbales sont respectivement désignées par les vocables « entités » et « épreuves » ; cette dernière notion, issue des nouvelles sociologies (Corcuff, 1995), désigne toute possibilité d'un changement d'état (Latour, 1984 ; Chateauraynaud, 1991). Parmi les fonctionnalités disponibles, je vais me focaliser sur le réseau, opération principalement assurée par Caliban mais également mobilisée par Marlowe.

4.1. *Des réseaux de cooccurrences*

Sous Caliban, l'algorithme du réseau travaille sur les listes d'entités. Il mesure le degré d'association de couples d'entités au sein d'un corpus. Cette mesure d'association est calculée comme suit : lorsque deux entités apparaissent dans la même phrase (c'est-à-dire dans la même chaîne de caractères située entre deux points), il y a cooccurrence. Si, parmi les mots qui se trouvent entre les deux entités se trouve au plus une épreuve, la valence de l'association est fixée à deux. S'il y a plus d'une épreuve entre les deux entités, la valence est de un pour cette phrase[8]. L'association entre deux termes peut-être jaugée pour un texte (on parle alors d'un réseau local) ou tout le corpus (il s'agit du réseau global) : les scores de chaque phrase sont alors simplement additionnés pour donner la mesure qui est utilisée dans la fenêtre des réseaux. Les profils d'associations d'une entité pour différents textes (ie les différents réseaux locaux) peuvent être comparés au réseau global, ce qui permet d'évaluer les divergences et originalités des différents textes.

La fenêtre du réseau global (voir ci-dessous) indique, pour un terme, l'ensemble des entités qui lui sont associées (et la force de ces associations). Le réseau est affiché en double : la

[7] Le lecteur désireux d'en savoir plus peut consulter le site http://www.prosperologie.org/.

[8] Le recours à cette valence permet de tenir compte de la « proximité » sur l'axe syntagmatique sans restreindre le co-texte au simple entourage du terme.

colonne de gauche indique le réseau d'entités, mais agrège les scores des entités contenues dans des êtres fictifs ; alors que la colonne de droite n'indique que les scores agrégés sous les catégories d'entités (les notions d'« êtres fictifs » et de « catégories » sont définis dans les points suivants).

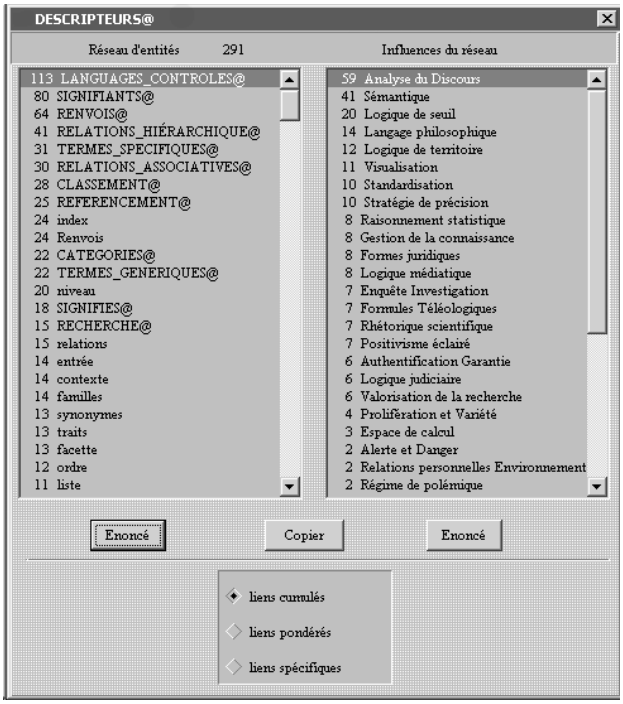


Fig. 3. Réseau global de l'entité DESCRIPTEURS@

Le réseau illustré indique l'univers conceptuel dans lequel est plongée une entité (Chateauraynaud 2003) ; il permet de suivre les traductions opérées par les différents acteurs (Callon, 1986). Ces traductions renseignent sur les connexions opérées par les acteurs entre différents thèmes ou préoccupations (Callon, 1993 ; Callon *et al.*, 1991b). Au bas de la fenêtre illustrée figurent deux autres options de calcul. Tout d'abord, l'algorithme dit des « liens pondérés » modifie la force des liens en fonction de la proportion de textes dans lesquels l'association est présente. Le calcul des « liens spécifiques » débarrasse quant à lui le réseau de la tête de liste d'entités associées ; celles-ci étant centrales dans le dossier, elles sont associées à tous les termes, ce qui est par conséquent peu parlant. Les liens spécifiques augmentent donc la probabilité de représenter des liens spécifiques les plus pertinents.

L'examen de la fenêtre du réseau global amène à trois remarques (que je commente dans les points suivants) : (1) aucune représentation graphique n'est ici mobilisée, contrairement à Candide, Leximappe, Réseau-Lu ou T-Lab ; (2) les réseaux sont parsemés de constructions de l'utilisateur (contrairement aux agrégats, qui sont générés par algorithme) ; (3) les deux colonnes représentent de manière différente le même réseau, en recourrant pour la première à des « ETRES-FICTIFS@ » et, pour la seconde, à des « Catégories ».

4.2. *Pas de graphe*

Contrairement au rendu des résultats exposés jusqu'ici, les concepteurs ont délibérément opté pour une présentation en liste. Ce choix a été motivé par les propriétés de ce format de présentation. Contrairement aux graphes réticulaires, en effet, la liste ne change pas de nature

lorsque son volume augmente. Ceci permet d'éviter les problèmes de lecture des réseaux touffus que le projet Textarc, par exemple, donne à voir (http://textarc.org/).

Fig. 4. Carte TextArc

Ce projet ne poursuit pas de retour critique sur la représentation des cartes de liens. Néanmoins, n'introduisant ni seuil ni pondération dans sa construction, il aboutit à des résultats qui ne gagnent ni en synthèse ni en clarté. Le travail avec des listes permet également de raisonner sur la queue de la série, donc de baser des interprétations sur les associations minoritaires, chose que l'accumulation quantitative des cartes cognitives interdit par définition (Chateauraynaud, 2003).

4.3. *Un masque de descripteurs surajoutés*

Outre l'absence de génération cartographique, les principe de construction des "agrégats" change également. Là où, précédemment, la composition des "clusters" était automatisée, elle est ici laissée à l'appréciation du chercheur. Cette différence influence donc les résultats et, par conséquent, la façon dont il faut les lire. Dans le premier cas, la cooccurrence amenait à agréger les liaisons les plus fortes représentant, selon les défenseurs de ce type d'outil (Callon, 1993), les thèmes, problèmes ou questions de préoccupation les mieux constitués. Dans le cas de Prospéro, l'accumulation des scores de plusieurs mots liés se fait selon d'autres principes que la connexion. Je vais les détailler et en dégager les conséquences.

La fenêtre comporte les constituants du réseau d'uns seule entité (ce dernier apparaît dans la barre de titre plus foncée en haut de la ***Fig. 3***). Outre les mots associés à ce terme (qui apparaissent en minuscules dans la première colonne) figurent des constructions de l'analyste qui

cumulent l'indice d'association de plusieurs entités. Dans la première colonne, ces constructions sont des êtres fictifs (leur nom apparaît en majuscules suivi d'un arobase @) ; dans la seconde, il s'agit de catégories d'entités (dont le nom est capitalisé).

4.3.1. *Êtres fictifs*

Les êtres fictifs réunissent la plage de variation des désignations d'un même actant sous une même fiction. Ces constructions – réalisées par le chercheur – consistent donc en une inscription des équivalences : l'opération consiste à collecter l'ensemble des désignations d'un même référent. La nature des objets étudiés en nouvelle sociologie (les controverses, les enrôlements, les disputes, etc.) rend ce travail d'autant plus nécessaire qu'ils impliquent plusieurs acteurs (dont les points de vue sont concurrents). Les variations de désignations se multiplient donc d'autant.

Comme le précise Sacks, cette multiplication des dénominations d'un même référent par les acteurs témoigne non seulement de leur intérêt pour celui-ci, mais également d'une « administration » du domaine qu'il recouvre. Cette gestion terminologique met en scène l'exercice d'une autorité : tenter d'imposer un nom revient à exercer un « contrôle social » ; le contester est une forme de « rébellion » contre celui-ci (Sacks, 1992a ; Silverman, 1998).

La plage des variations terminologiques qui composent un être fictif procède de la co-référence ou de la synonymie. La synonymie est entendue dans le sens suivant : suivant un point de vue de sociologie pragmatique, sont considérés synonymes deux termes que les acteurs mobilisent pour désigner le même phénomène. La construction des êtres fictifs n'est cependant pas indépendante du contexte : selon les affaires, les acteurs distinguent des désignations qui, dans d'autres circonstances, sont assimilées. Tantôt policiers et gendarmes seront substituables en tant que représentants des forces de l'ordre, tantôt ils se distribuent selon deux groupes aux intérêts divergeant (ce qui fut en partie le cas, par exemple, lors de la réforme belge des polices). Les êtres fictifs sont donc construits en fonction de l'étude sociologique.

4.3.2. *Catégories*

Dans la seconde colonne, on trouve des catégories. Contrairement à la colonne des êtres-fictifs, aucun terme n'apparaît seul. C'est pourtant exactement le même réseau que celui de la première colonne qui est représenté ici. Mais figurent uniquement les scores des termes codés sous une des catégories du cadre d'analyse du sociologue. Celles-ci ne sont plus construites comme une variation de la désignation. Les catégories recouvrent plutôt des univers de cohérence qui prennent sens dans un corpus déterminé et en fonction de l'analyse. Elles permettent d'asseoir l'interprétation sociologique et sont par conséquent modifiées, fusionnées ou disséquées en fonction du sujet traité, du contexte ou de ce que l'on veut montrer. Une catégorie correspond à une logique sociale, un registre d'action, un champ social (Bourdieu, 1980), une cité ou un monde commun (Boltanski et Thévenot, 1987).

4.4. *L'évolution du réseau*

Jusqu'ici, j'ai détaillé le calcul du réseau de manière statique : ce dernier agrège les scores des associations de tous les textes traités et synthétise donc ces associations sans tenir compte de la nature diachronique du corpus (la temporalité est donc écrasée). Or, les corpus étudiés en nouvelle sociologie sont souvent des affaires impliquant non seulement de multiples acteurs mais de nombreux événements. Les associations dont le réseau rend compte sont donc mouvantes au cours du déroulement du dossier. Pour rendre compte de cette dynamique temporelle, l'évolution et les reconfigurations successives du réseau doivent être examinées. On

peut ainsi identifier des périodes sur base de la stabilité des associations (ou, pour l'exprimer autrement, des dates de rupture, caractérisées par une modification des têtes de réseau).

4.5. *Le retour réflexif assisté par ordinateur*

Enfin, le réseau permet de tester la pertinence du codage des catégories. Cette pertinence du cadre d'analyse est tributaire de son adéquation aux données dont il entend rendre compte. Aussi le réseau fournit-il une ressource simple pour éprouver la pertinence du codage de l'utilisateur (c'est-à-dire du sociologue). La façon de procéder est la suivante. L'algorithme sélectionne parmi les termes appartenant à une catégorie celui qui établit le meilleur score (la plus grande occurrence). Le réseau de mots associés à cette forme nominale est examiné. Si, dans la tête de celui-ci, la catégorie sous laquelle il est codé est représentée, alors ce recouvrement est considéré comme un bon indice de la pertinence de la catégorie en question pour représenter les enjeux du corpus en présence.

5. Conclusions

Du côté de l'automatisation, j'apprécie le large éventail de résultats déductibles des réseaux de mots associés, qu'il s'agisse de ceux de Leximappe, de Candide, de Réseau-Lu, de T-Lab ou de Prospéro. Ce dernier logiciel est moins attractif que les autres : du coté de l'apparence, les listes ont un côté spartiate un peu rebutant. Ceci n'est pas qu'un jugement esthétique : la présentation des résultats en liste semble intéresser bien moins que les graphes (qu'il s'agisse des comités de revues scientifiques ou de l'assistance des colloques).

Néanmoins, la voie choisie par les concepteurs de ce dernier logiciel m'a séduit. Contrairement aux agrégats générés automatiquement, des constructions comme les êtres-fictifs et les catégories permettent au chercheur d'assurer la solution de continuité entre méthode et orientation théorique. Par conséquent, les outils offerts par Prospéro n'enferment pas leurs utilisateurs dans un paradigme scientifique déterminé (Kuhn, 1983) : ceux-ci peuvent tout aussi bien être psychanalystes, bourdieusiens ou ethnométhodologues. Prospéro consiste donc en un bon outil résistant aux boîtes noires. Mais cette liberté a comme pendant une responsabilité : chaque chercheur est responsable de ses catégories d'analyse, donc de son interprétation. Cet outil rend à l'activité scientifique sa dimension spécifique telle qu'elle a été analysée par les philosophes des sciences : la prise de risque (Stengers, 1992).

Références

Akrich M. (1991). L'analyse socio-technique. In Vinck D. (sous la coordination de), *Gestion de la recherche. Nouveaux problèmes, nouveaux outils*. Armand Colin.

Boltanski L. et Thévenot L. (1987). *Les économies de la grandeur*. PUF.

Boltanski L. et Chiapello E. (1999). *Le nouvel esprit du capitalisme.* Gallimard.

Bourdieu P. (1980). *Le sens pratique*. Minuit.

Bourdieu P., Chamboredon J.-C. et Passeron J.-C. (1968). *Le métier de sociologue*. Mouton/Bordas.

Bourdoncle F. (1997). LiveTopics : Recherche Visuelle d'Information sur l'Internet. *La documentation française, Dossiers de l'audiovisuel*, vol. (74) : 36-38.

Callon M. (1986). Éléments pour une sociologie de la traduction. La domestication des coquilles Saint-Jacques et des marins-pêcheurs dans la baie de Saint-Brieuc. *L'Année sociologique*, vol. (36) : 169-208.

Callon M. (1993). *La scientométrie*. PUF.

Callon M., Courtial J.-P. et Turner W. (1991a). La méthode Leximappe : un outil pour l'analyse stratégique du développement scientifique et technique. In Vinck D. (sous la coordination de), *Gestion de la recherche. Nouveaux problèmes, nouveaux outils*. Armand Colin.

Callon M., Courtial J.-P., Turner W. et Bauin S. (1991b). From translations to problematic networks : An introduction to co-word analysis. *Information sur les sciences sociales*, 2 (vol. 22) : 191-235.

Chateauraynaud F. (1991). *La Faute professionnelle. Une sociologie des conflits de responsabilité*. Métailié.

Chateauraynaud F. et Torny D. (1999). *Les sombres précurseurs. Une sociologie pragmatique de l'alerte et du risque*. EHESS.

Chateauraynaud F. (2003). *Prospéro : Une technologie littéraire pour les sciences humaines*. CNRS.

Corcuff P. (1995). *Les nouvelles sociologies. Constructions de la réalité sociale*. Nathan.

Courtial J.-P. et Kerneur L. (1995). Contribution de l'analyse des mots associés au suivi du développement d'un champ scientifique. In Cocaud M. (textes réunis par), *Histoire et Informatique. Base de données, recherche documentaire multimédia*. PUR.

Conein B. (1993). L'ethnométhodologie comme sociologie descriptive : réflexivité et sui-référentialité des catégories sociales. *Cahiers de recherche ethnométhodologique*, vol. (1) : 73-88.

Deloison J.-P. (1994). *Acquisition automatique de connaissances linguistiques pour l'indexation automatique. Rapport de DEA « Contrôle des Systèmes »*. Université de Technologie de Compiègne.

Ducrot O. et Schaeffer J.-M. (1995). *Nouveau dictionnaire encyclopédique des sciences du langage*. Seuil.

Flake G., Lawrence S., Giles L. et Coetzee F. (2002). Self-Organization and Identification of Web Communities. *IEEE Computer*, vol. (35/3) : 66-71.

Garfinkel H. et Sacks H. (1970). On Formal Structures of Practical Actions. In McKinney J. et Tiryakian E. (Eds), *Theorical Sociology. Perspectives and Developments*. Appleton-Century-Crofts.

Jenny J. (1996). Analyse de contenu et de discours dans la recherche sociologique française : pratiques micro-informatiques actuelles et potentielles. *Current Sociology*, vol. (44/3) : 279-290.

Jenny J. (1997). Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. État des lieux et essai de classification. *Bulletin de Méthodologie Sociologique*, vol. (54) : 64-112.

Kuhn T. (1983). *La Structure des Révolutions Scientifiques*. Flammarion.

Latour Br. (1984). *Les microbes. Guerre et paix suivi de Irréductions*. Métailié/Pandore.

Latour Br. (1995). *La Science en action*. Gallimard.

Lejeune Chr. (2001). Du mode de définition de deux programmes de recherche en sociologie et en ethnométhodologie. *Carnets de bord*, vol. (2) : 56-66.

Lejeune Chr. (2002). Indexation et organisation de la connaissance. La régulation des décisions sur un forum de discussion. *Les cahiers du numérique*, vol (3/2) : 129-144.

Malingre M.-L. (1995). Une application de Candide, logiciel d'analyse textuelle pour une histoire de la traduction littéraire. In Cocaud M. (textes réunis par), *Histoire et Informatique. Base de données, recherche documentaire multimédia*. PUR.

Noyer J.-M. (1995b). Utilisation d'un outil Infométrique, « Candide » dans le contexte d'une réflexion stratégique. Les réseaux de simulation distribuée de l'armée américaine : émergence et description de l'émergence. In Noyer J-M. (sous la direction de), *Les sciences de l'information. Bibliométrie, scientométrie, infométrie*. PUR.

Osgood C. (1959). The representational model and relevant research methods. In De Sola Pool I. (Ed.), *Trends in Content Analysis*. University of Illinois Press.

Ramaux N. (1993). *Acquisition automatique et polysémique en Langage naturel. Rapport de D.E.A. « Contrôle des systèmes »*. Université de Compiègne.

Ramognino N. (1999). Linguistique et sociologie, un point de vue méthodologique. *Sociologie et sociétés*, vol. (XXXI/1) : 35-50.

Sacks H. (1992a). *Lectures on Conversation. Volume 1*. Blackwell.

Silverman D. (1998). *Harvey Sacks. Social Science & Conversation Analysis*. Polity Press.

Stengers I. (1992). *La volonté de faire science*. Les Empêcheurs de penser en rond.

Stone P., Dunphy D., Smith M. et Ogilvie D. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* The MIT Press.

Teil G. (1991). Candide™ : un outil de veille technologique basé sur l'analyse des réseaux. In Vinck D.(sous la coordination de), *Gestion de la recherche. Nouveaux problèmes, nouveaux outils*. Armand Colin.

Teil G. (1994). L'analyse des goûts alimentaires par les réseaux. Une cartographie des critères du goût des consommateurs. In Courtial J.-P. (sous la direction de), *Science cognitive et sociologie des sciences*. PUF.

Teil G. et Latour Br. (1995). The Hume machine. Can association networks do more than formal rules ? *Stanford Humanities Review*, vol. (4, issue 2).

Thomas A. et Shearer J. (2000). *Internet Searching and Indexing. The Subject Approach*. The Haworth Press.

Weitzman E. et Miles M. (1995). *A Software Source Book. Computer Programs for Qualitative Data Analysis*. Sage Publications.

Yates S. (1996). Oral and Written Linguistic Aspects of Computer Conferencing. A Corpus Based Study. In Herring S. (Ed.), *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*. John Benjamins.

Yates S. (2001). Researching Internet Interaction: Sociolinguistics and Corpus Analysis. In Wetherell M. Taylor S. et Yates S., *Discourse as Data. A Guide for Analysis*. The Open University.

Analyse en composantes locales et graphes de similarité entre textes

Alain Lelu

INRA / unité Mathématiques, Informatique, Génome
et Université de Franche-Comté / LASELDI
alain.lelu@univ-fcomte.fr

Abstract

Cosines between line-vectors (or column-vectors) of a frequency [texts × terms] matrix define 2 valued graphs. Using distributional (or "Hellinger") distance, transition formulas allow one to deduce important properties of a graph from the other's : the first eigenvalue of each cosine adjacency matrix can be interpreted as a centrality index for texts and terms ; the following two factors provide an optimal 2D mapping for the graph, as correspondence analysis does. Our Local Components Analysis method uses these properties 1) globally: the extended concept of K-reciprocal neighbors defines a graph with adaptive density indices for its nodes, allowing for Trémolière's percolation algorithm to cluster it, 2) locally: each cluster is characterized by its first eigenvalues (resulting in texts and terms centrality indices in the cluster, and projections of the whole corpus'texts and terms on the cluster axoïd), and the following two ones (2D mapping of the K-reciprocal links). Illustration is given with Corneille and Molière theater corpus.

Résumé

Les matrices de cosinus entre vecteurs-ligne et colonne d'un tableau de fréquences [textes × termes] définissent des graphes valués sur ces deux ensembles. L'utilisation de la distance distributionnelle (dite aussi de Hellinger) fournit des relations de transition entre propriétés importantes calculées sur ces deux graphes : le 1er vecteur propre de chaque matrice d'adjacence fournit les indices de centralité des textes (resp. mots), les 2 suivants, proches des 2 premiers facteurs non triviaux d'AFC, permettent une représentation optimale plane des graphes. L'Analyse en Composantes Locales utilise ces propriétés : 1) au niveau global, la notion de voisinage K-réciproque détermine un graphe de contiguïté où chaque nœud est doté d'une valeur de densité *adaptative*, utilisée par l'algorithme de percolation de Trémolières pour déterminer des classes denses, 2) au niveau local chaque classe est caractérisée par ses 1ers vecteurs propres (définis comme ci-dessus : centralités des textes et des termes dans chaque classe) et les projections de tous les termes et textes du corpus sur l'axe des classes. Et par ses deux suivants (en tant que fond de carte pour représenter ses graphes de liens K-réciproques). Ces points sont illustrés à partir des pièces de Corneille et Molière indexées par D. Labbé.

Mots-clés : graphes, voisins réciproques, distance de Hellinger, densité, centralité, classification automatique, percolation, analyse en composantes locales.

1. Introduction

Les méthodes de fouille de textes, objets d'importants enjeux (génomique fonctionnelle, veille scientifique…) se heurtent à deux difficultés : 1) celle de la sélection des descripteurs (ici : termes lemmatisés ou non, composés ou non) pertinents, 2) celle de l'unicité de la représentation obtenue : seules des méthodes convergeant vers un maximum absolu d'un critère de qualité peuvent rendre compte de la variation dans le temps des représentations ; et il est toujours souhaitable pour l'expert qui les interprète de disposer d'une représentation stable et unique pour une collection donnée de textes.

De nombreux travaux récents sur les graphes à invariance d'échelle (scale-free) et dont la signature est spécifique (répartitions à lois de puissance, comme la « loi de Zipf » dans le domaine textuel), renouvellent le point de vue sur le domaine (Watts et Strogatz, 1998). Nous présentons ici une formulation en termes de graphes de nos travaux sur la représentation dynamique d'un flux de textes (Lelu et Ferhan, 1998 ; Lelu et François, 2003).

2. Constituer les graphes de similarité entre textes :

2.1. *Sélectionner les descripteurs pertinents :*

Disposer d'un bon extracteur de lemmes et termes composés n'est plus aujourd'hui une gageure. Mais la quantité de candidats termes obtenue, autant que le caractère « Zipfien » de la répartition des mots (il existe un continuum allant des mots courants non contextuels aux nombreux mots de fréquence faible, en passant par les mots de fréquence moyenne caractérisant tel ou tel contexte sémantique du corpus) exigent un processus d'élimination des deux types extrêmes de termes. Aujourd'hui nécessairement supervisé par un expert, ce processus doit être intégré dans une boucle itérative incluant une évaluation (automatique dans la mesure du possible) de la qualité de la représentation obtenue. Ce qui constitue un problème en soi, dont nous ne parlerons pas ici ; c'est sans aucune sélection de mots que nous illustrerons notre propos, sur le corpus des pièces de Corneille et Molière indexé par Dominique Labbé (2001) et mis à disposition de la communauté scientifique par celui-ci (67 œuvres caractérisées par 9946 termes grammaticalement typés, dont 3076 hapax, pour un total de 928 000 occurrences).

2.2. *Choix de l'indice de similarité*

Dans notre domaine d'application, plusieurs critères doivent caractériser cet indice :

– Ne pas dépendre de la taille des textes : les vecteurs caractérisant les textes doivent être normalisés.

– Ne pas dépendre de l'arbitraire du choix des mots par les auteurs quand ils expriment un certain contenu sémantique – souvent en référence à des critères esthétiques comme éviter de répéter le même mot à peu de distance : ce qui se traduit par le critère d'« équivalence distributionnelle » (stabilité des similarités par rapport à la fusion ou à l'éclatement de descripteurs de répartitions voisines), satisfait, entre autres, par la transformation de l'espace des données dans lequel opère l'analyse factorielle des correspondances (AFC).

Notre réponse à cette double exigence est remplie par la distance distributionnelle d(t,t') (dite aussi de Hellinger (Domengès et Volle, 1979)), évaluée dans (Lelu, 2003), et utilisée par ailleurs (Bacelar-Nicolau, 1987) : $\mathrm{Sim}(t, t') = 2\text{-}d(t, t')^2 = \cos(x_t, x_{t'}) = \langle \underline{x}_t, \underline{x}_{t'} \rangle$

où $\underline{x}_t = \{\ldots \sqrt{(x_{it}/x_{.t})} \ldots\}$ est un vecteur-texte normalisé, x_{it} est la fréquence du mot i dans le texte t, $x_{.t}$ est la fréquence totale des mots du texte t, et <..., …> symbolise le produit scalaire.

2.3. *Du tableau des similarités au graphe de similarité*

Le tableau complet des similarités entre les textes du corpus peut être considéré comme la matrice d'adjacence d'un graphe valué porteur de la totalité de l'information disponible sur les similarités. Cependant, l'examen direct de celui-ci, à diverses valeurs de seuil de similarité, par des outils d'exploration de graphe (par ex. BioLayout (Enright et Ouzounis, 2001)) ne fait ressortir le plus souvent qu'une grande classe connexe et touffue entourée par quelques classes de faibles effectifs et beaucoup de points isolés, du fait de la répartition très inégalitaire, Zipfienne, des mots.

3. Exploiter les graphes de similarité entre textes :

Il est donc impératif de styliser ce graphe en « augmentant le contraste » de façon pertinente, si l'on veut isoler des cliques de documents thématiquement homogènes. Plusieurs méthodes ont été proposées pour ce problème de « graph partitioning » (pour une revue, cf. (Brandes *et al.*, 2003)), qui font le plus souvent appel à des processus itératifs.

Pour notre part nous avons opté pour une méthode utilisant une mesure de densité locale des nœuds d'un graphe, inspirée de l'algorithme de percolation de (Trémolières, 1979) : à paramètre de finesse d'analyse fixé, l'énumération des maxima locaux de densité constitue l'optimum global recherché. On obtient ainsi 1) un ensemble de noyaux de classes, 2) un ensemble de nœuds ambivalents, communs à plusieurs noyaux, 3) les nœuds isolés restants peuvent être projetés sur les axes caractérisant les noyaux (cf. §3.1) autorisant ainsi une notion de classes floues et recouvrantes, plus « naturelle » dans notre domaine d'application que celle de partition stricte. À noter que (Colombo *et al.* ,2003) utilisent le même principe de recherche de maxima de densité, avec un principe d'extension de ceux-ci moins riche que dans la méthode de Trémolières.

Reste le choix de la fonction densité ; après des essais décevants avec un noyau de taille fixe, paramétrable (Lelu, 1994), dus aux disparités considérables de densité des classes dans un espace à grand nombre de dimensions, nous avons adopté la mesure *adaptative* de densité suivante : 1) parmi les K plus proches voisins de chaque texte, définis (sans garantie de symétrie) sur la matrice de similarité, nous ne retenons que les voisins réciproques, 2) la densité de chaque nœud est une version modifiée du « clustering coefficient » (Albert et Barabasi, 2001) : 1 (si le noeud a un ou plusieurs 1-voisins) + (nombre d'arêtes de son 1-voisinage non rattachées au nœud considéré / nombre maximum possible de ces arêtes) ; ainsi un point isolé a une densité de 0, le centre d'une « étoile », ou chaque élément d'un couple isolé ont une densité de 1, et toute autre configuration une densité comprise entre 1 et 2. Le graphe correspondant traduit l'« essence » du graphe complet, le stylise de façon robuste : les voisins réciproques marquent l'extremum d'un gradient de densité croissante, le long duquel les points se rapprochent de plus en plus, ou des « grumeaux » accidentels intéressants en soi à repérer, au même titre que les points isolés. Cette notion de K-réciprocité étend celle de voisins réciproques (1-réciprocité) déjà utilisée en classification automatique (Benzécri, 1982).

La figure 1 montre les voisinages 3-réciproques des 67 œuvres de Corneille et Molière dans le plan des facteurs 2 et 3 d'analyse factorielle sphérique du corpus présenté (où les positions relatives des œuvres sont les mêmes, à quelque % près, que celles obtenues par AFC). Les deux croix dans la partie négative de F1 sont *Le Menteur* et *La suite du Menteur* , comédies tardives de Corneille ; le rond dans la partie positive de F1 est *Dom Garcie de Navarre* (seule tentative de comédie héroïque de Molière).

4. Caractériser chaque noyau de textes

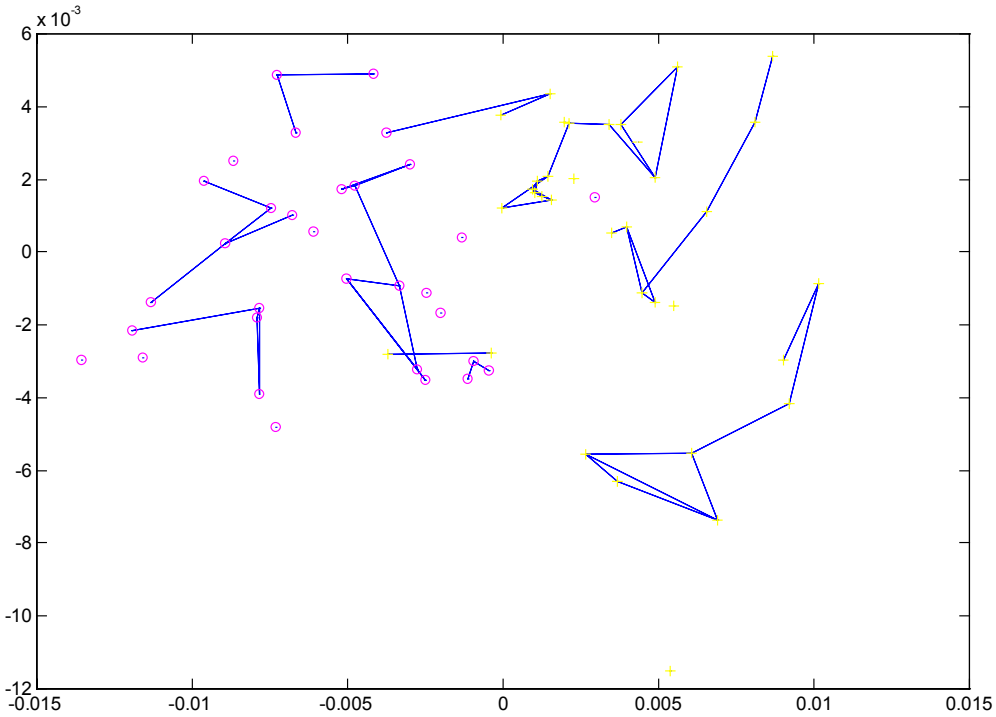
Le choix de l'espace distributionnel est également pertinent au niveau local de chaque noyau :

4.1. *Centralités des textes, des mots, et de variables explicatives extérieures*

Le premier vecteur propre de la matrice de similarité entre textes du même noyau (pondérés par leur nombre total de mots) constitue un indicateur de centralité de chaque texte au sein de son noyau (cf. Brandes, 2003) pour une revue sur l'*eigenvector centrality*) ; la pondération choisie permet d'en déduire l'indicateur de centralité de chaque mot au sein du graphe

« dual » des similarités locales entre mots de ce noyau (Domengès et Volle, 1979). La projection de l'ensemble des textes et des mots du corpus sur l'axe du noyau considéré illustre la « zone d'influence » de ce noyau : certains textes, isolés ou appartenant à d'autres noyaux, peuvent se projeter « haut », s'ils comportent plus d'un thème abordé. Nous verrons dans la section suivante l'illustration de ces principes pour caractériser des sous-ensembles homogènes de pièces dégagés par notre méthode.

Cette même procédure permet également de situer l'importance explicative de variables extérieures à l'analyse comme l'auteur du texte, la source, l'année de parution, etc. en les projetant à titre illustratif sur l'axe du noyau.



Graphe des voisins 3-réciproques sur fond (plan F2 × F3) d'Analyse Factorielle Sphérique (Corneille : + ; Molière : o).

4.2. *Visualisation des graphes locaux*

Les mêmes auteurs ont montré que les projections des textes sur les vecteurs propres suivants constituent une approximation des facteurs non triviaux extraits par AFC de la matrice des fréquences brutes. Or des travaux récents (Brandes, 2003), qui rejoignent une lignée plus ancienne (Lebart, 1984), montrent que les avant-derniers vecteurs propres de la matrice « laplacienne » issue de la matrice d'adjacence, autrement dit les premiers facteurs non triviaux d'AFC sur cette matrice d'adjacence, permettent une visualisation graphique optimale dans deux (ou trois) dimensions, optimale au sens d'un *spring-embedder algorithm* particulier : celui d'un système d'anneaux (nœuds) reliés par des ressorts (arcs) de longueur nulle au repos dans un espace à 2 (ou 3) dimensions. On peut également en déduire la représentation dans 2 ou 3 dimensions du graphe « dual » des mots (cf. §3).

5. Résultats et perspectives

L'ACL de notre corpus extrait 23 pièces isolées (en majorité des comédies en prose de Molière) et 9 noyaux de classes :

– 3 noyaux : comédies en prose de Molière
– 1 noyau : comédies en vers de Molière
– 1 noyau : comédies et tragi-comédies de Corneille (en vers)
– 3 noyaux : tragédies de Corneille (en vers)
– 1 noyau : les 2 versions de *Psyché* (co-écrite par Corneille et Molière) présentes dans le corpus.

Aucun de ces noyaux ne mélange les œuvres des deux auteurs. Les deux seules proximités notables indiquées par les projections sont celle du *Menteur* et de la *Suite du Menteur* avec les grandes comédies en vers de Molière, et de *Psyché* avec le noyau *Amants magnifiques* et *Princesse d'Elide*.

Chaque noyau est caractérisé par ses textes et mots les plus centraux (cf. ci-dessous le noyau le plus dense : les grandes comédies en prose de Molière).

Noyau N°1 :

Densité :	*Centralité :*	*N° int. :*	*Titre (v : vers ; C : comédie) :*
2.00	(.947)	64 =	Les_Fourberies_de_Scapin___C
2.00	(.957)	51 =	Dom_Juan___C
1.50	(.945)	59 =	George_Dandin__C
1.40	(.965)	60 =	L'Avare___C+
1.00	(.943)	63 =	Le_Bourgeois_gentilhomme__C
Individus ambivalents :			
1.40	(.950)	45 =	L'Ecole des femmes_v_C
1.30	(.953)	50 =	Le_Tartuffe_v_C
1.00	(.942)	67 =	Le_Malade imaginaire___C
Individus extérieurs les plus proches :			
1.30	(.935)	39 =	Dépit_amoureux_v_C
1.33	(.932)	38 =	L'Etourdi_v_C
2.00	(.932)	53 =	Le_Misanthrope_v_C

Mots :

Centralité :	*Freq.locale :*	*Freq.relative (‰) :*	*Libelle/code grammatical:*
0.9997	6077	52.5	de/8_____
0.9993	2731	23.6	que/8_____
0.9992	2453	21.1	ne/6_____
0.9992	5895	50.9	le/7_____
0.9991	2317	20.0	il/5_____
0.9991	1955	16.8	un/7_____
0.9990	95	0.82	foi/2_____
0.9990	2407	20.8	à/8_____
0.9990	693	5.98	voir/1_____
0.9990	1403	12.1	faire/1_____
0.9989	3631	31.3	être/1_____
0.9987	1734	14.9	que/5_____

Les mots les plus centraux du noyau sont ceux qui présentent les fréquences relatives les plus constantes dans les trois œuvres. Pour interpréter les thèmes en matière de style et de contenu, ceci peut fournir un angle de vue complémentaire de ceux donnés par d'autres indicateurs, comme l'écart réduit à la fréquence moyenne dans le corpus – ainsi *peu* et *fort* sont les mots les plus centraux de trois grandes comédies en vers de Molière, alors qu'*avoir*, *à*, *faire* le sont dans trois tragédies de Corneille. Une grande finesse d'analyse est ainsi possible à partir de notre extraction d'œuvres isolées et de noyaux denses : on peut les placer sur un graphe d'ensemble, les caractériser par plusieurs indices et listes ordonnées de mots, les représenter

localement sur un axe de centralité, aussi bien que sur un plan factoriel local si ce noyau est complexe. Ces possibilités nous paraissent intéressantes si elles sont appropriées par des experts du domaine qui peuvent les mettre en relation avec leur connaissance fine des textes et des éléments d'analyse extérieurs à ceux-ci ; moins intéressantes en tant que « tranchoir » pour un usage décisionnel d'attribution d'auteur…

Au delà d'une application « grandeur réelle » à 1400 enregistrements bibliographiques médicaux (Lelu et François, 2003), qui a donné lieu à un graphe de 232 noyaux consultable sur une interface Web graphique, notre horizon est celui de l'analyse dynamique de corpus : cette méthode produit une séquence réticulée de classes quand on introduit de nouveaux textes dans l'analyse ; chaque texte, ou paquet de textes, nouveau entraîne des créations /suppressions/ fusions /éclatements localisés de noyaux existants. La difficulté est alors de concevoir et réaliser une interface conviviale pour parcourir dans ses dimensions tant spatiales que temporelle ce réticulogramme de construction des classes.

Bibliographie

Bacelar-Nicolau H. (1987). On the distribution equivalence in cluster analysis. In Devijver PA and Kittler J. (Eds), *Patt. Rec. Theory and Appl.*, NATO ASI series F, vol. (30), Springer-Verlag : 73-79

Benzécri J.P. (1982). Construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Données*, vol. (2) : 208-218.

Brandes U., Gaertler M. et Wagner D. (2003). Experiments on Graph Clustering Algorithms. In Proceedings of the *11th Europ. Symp. Algorithms* (ESA '03). Springer LNCS.

Colombo T., Quentin Y. et Guénoche A. (2003). Recherche de zones denses dans un graphe : application aux gènes orthologues. In San Juan E. (Ed.), *JIM'2003*, sept. 3-6, 2003, Metz.

Domengès D. et Volle M. (1979). Analyse factorielle sphérique : une exploration. *Annales de l'INSEE*, vol. (35) : pp. 3-84.

Enright AJ et Ouzounis C. (2001). BioLayout--an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, vol. (17/9) : 853-4.

Labbé D. et Labbé C. (2001). Inter-Textual Distance and Authorship Attribution. Corneille and Molière. *Journal of Quantitative Linguistics*, vol (8/3) : 213-231.

Lebart L. (1984). Correspondence Analysis of graph structures. *Bull. Techn. du CESIA*, vol. (2/1-2) : 5-19.

Lelu A. (1994). Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In Diday E., Lechevallier Y. *et al.* (Eds), *New Approaches in Classification and Data Analysis*. Springer-Verlag : 241-248.

Lelu A. et Ferhan S. (1998). Clustering a textual dataflow by incremental density-modes seeking. In Proceedings of *IFCS'98* (International Federation of Classification Societies) : 206-209.

Lelu A. (2003). Évaluation de trois mesures de similarité utilisées en sciences de l'information. Information Sciences for Decision Making, vol. (6).

Lelu A. et François C. (2003). Un algorithme de détection de maximas de densité basé sur la distance distributionnelle : application à la classification optimale fine d'un corpus documentaire. In Dodge Y. et Melfi G. (Eds), *10èmes Rencontres SFC*, Pr. Acad. Neufchâtel.

Trémolières R.C. (1979). The percolation method for an efficient grouping of data. *Pattern Recognition*, vol. (11/4).

Watts D. et Strogatz S.H. (1998). Collective dynamics of 'small-world' networks. *Nature*, vol. (393) : 440-442.

Temps verbaux, axe syntagmatique, topologie textuelle : analyses d'un corpus lemmatisé

Dominique Longrée[1], Xuan Luong[2], Sylvie Mellet[3]

[1]« Bases, Corpus et Langage » (UMR 6039), Univ. d'Angers – 49000 Angers – France
[2]« Bases, Corpus et Langage » (UMR 6039), UNSA – B.P. 209, 06204 Nice cedex 3 – France
[3]« Bases, Corpus et Langage » (UMR 6039), CNRS – B.P. 209, 06204 Nice cedex 3 – France
longree@fusl.ac.be, mellet@unice.fr

Abstract

The intertextual distance is often calculated by studying frequencies. As part of a research project on texts topology, the methods described here try to take into account the text linearity in this calculation. In fact, we will study the distribution of verbal forms, according to their tenses, in the general framework of the text. The corpus is based on a latin historical lemmatized texts databank. Several methods will be compared : texts segmentation by division or by neighborhood, distances representation by curves or trees.

Résumé

La mesure des distances entre les textes est le plus souvent effectuée à partir du dénombrement des unités d'analyse. Dans le cadre d'une réflexion générale sur la topologie textuelle, nous présentons quelques outils développés pour prendre en compte, dans ce calcul, la linéarité du texte et les localisations des unités au fil de cette chaîne linéaire. En l'occurrence, il s'agit d'examiner la répartition globale des temps verbaux le long de l'axe syntagmatique pour évaluer les distances et similarités entre des textes lemmatisés d'historiens latins. Plusieurs méthodes sont comparées : segmentation des textes par découpage ou par voisinage, représentation des distances par courbes ou par analyses arborées.

Mots-clés : distance intertextuelle, topologie textuelle, mesures de voisinage, courbes de distribution, analyses arborées, temps verbaux, latin.

1. Objectifs

La recherche ici exposée représente une étape méthodologique importante dans le cours de recherches entamées depuis quelques années pour tenter d'améliorer les calculs de distance entre les textes en prenant en compte deux paramètres originaux : travaillant sur des textes latins lemmatisés et étiquetés[1] les auteurs s'intéressent en effet depuis longtemps d'une part au rôle des catégories grammaticales dans la caractérisation d'un texte, d'autre part à la distribution de ces catégories au fil des textes, c'est-à-dire à la topologie textuelle entendue au sens des mathématiciens (cf. Luong et Mellet, 1995 ; Salem, 2002 ; Lamalle et Salem, 2002 ; Longrée et Luong, 2003). En intégrant ces deux paramètres, très généralement délaissés dans les études de statistique linguistique, nous espérons pouvoir mieux évaluer les distances et similarités intertextuelles. Les catégories grammaticales retenues ici seront les temps et modes verbaux, paramètres jugés pertinents pour la comparaison des textes étudiés ; on espère ainsi obtenir un critère efficace (parmi d'autres sans doute) de reconnaissance et de classification

[1] Nos corpus reposent sur les fichiers informatisés et lemmatisés par le L.A.S.L.A. à l'Université de Liège. Pour cette étude, nous avons travaillé sur l'ensemble du corpus césarien, y compris ses continuateurs, sur l'ensemble des œuvres conservées de Quinte-Curce et de Tacite, et sur 9 *Vies* extraites des *Vies des douze Césars* de Suétone. Pour une présentation plus détaillée, cf. Longrée et Luong (2003).

par auteurs et, surtout, par sous-genres. L'objectif ultime est en effet de pouvoir caractériser de manière endogène et aussi objective que possible, au sein de l'histoire latine, le sous-genre du commentaire, celui des annales, celui des biographies, etc. Le propos sera, plus précisément, de présenter quelques outils que nous avons développés pour prendre en compte les localisations des unités d'analyse dans la chaîne linéaire des textes : segmentation des textes par découpage et voisinage, représentation des distances par courbes de distribution et analyses arborées.

2. L'emploi des temps verbaux

Le terme *emploi* recouvre au moins trois réalités, trois objets d'étude statistique différents.

2.1. ***Les fréquences*** : les unités d'analyses sont considérées globalement et en vrac ; elles donnent lieu à des dénombrements qui permettent de constituer de classiques tableaux de contingence à partir desquels des AFC, des analyses arborées ou des analyses en composantes principales permettront de représenter les similarités des profils textuels (cf. Longrée, à paraître 2004). Il est à noter cependant que les matrices issues du dénombrement des catégories grammaticales présentent des spécificités par rapport aux matrices lexicales qui nécessitent un aménagement des analyses traditionnelles (obligation de travailler sur les fréquences et non pas en termes de présence / absence ; prise en compte de quelques colonnes creuses et néanmoins significatives (cf. Luong et Mellet, 2003)).

2.2. ***Les séquences*** : elles intègrent l'axe syntagmatique en ce qu'elles rendent compte des effets de succession — enchaînements ou ruptures — dans l'emploi des temps verbaux. Certaines séquences sont linguistiquement conditionnées : ainsi, en latin, est-il rarissime qu'un infinitif de narration apparaisse isolément ; on a plutôt affaire à des séries d'au moins trois occurrences et souvent plus. La séquence des temps en propositions subordonnées est le plus souvent contrainte par le temps du verbe principal (phénomène dit de « concordance des temps »). Mais la plupart des enchaînements de proposition principale à proposition principale sont libres. C'est pourquoi nous avons réduit nos textes à la simple succession des codes caractérisant les formes verbales des propositions principales et indépendantes afin de pouvoir dénombrer les séquences de deux codes identiques successifs, puis de trois, de quatre, de cinq, etc. codes identiques successifs (soit de manière absolue, soit de manière cumulative, une séquence de quatre codes identiques successifs étant alors également comptabilisée comme 2 séquences de trois codes et 3 séquences de deux codes ; complémentairement, cette réduction permet aussi de mettre en évidence et d'analyser les types de ruptures dans la succession des temps principaux. Les résultats, publiés dans Longrée et Luong (2003) et Longrée (à paraître, 2003) sont prometteurs.

2.3. ***La répartition globale sur l'ensemble d'un texte*** : il s'agit d'une troisième réalité qui associe la saisie de l'axe syntagmatique à courte portée déjà présente dans les séquences et une saisie à longue portée qui tient compte de la répartition des unités étudiées dans les différentes zones du texte. C'est là qu'on retrouve véritablement la notion de topologie textuelle.

3. Les méthodes : les représentations graphiques

Il s'agit donc de synthétiser les informations contenues dans les deux graphes ci-dessous et de leur donner une forme exploitable pour un calcul de distance.

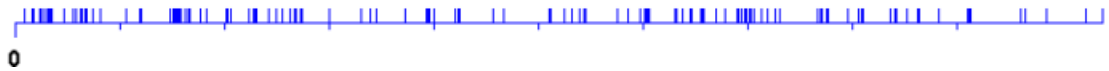


Figure 1. Répartition linéaire des imparfaits dans la Guerre d'Alexandrie

Ce graphe visualise la répartition des formes d'imparfaits dans la chaîne linéaire du texte : il montre assez clairement les zones de densité importante dans les premier et troisième quarts du texte, et les zones de faible présence dans les deux autres quarts, tout particulièrement à la fin du texte. Les données numériques associées à ce graphe sont les suivantes :

Nombre de mots : 10666

Nombre d'imparfaits : 160

Distance moyenne théorique entre deux imparfaits : 66,66

Distance maximale observée : 482

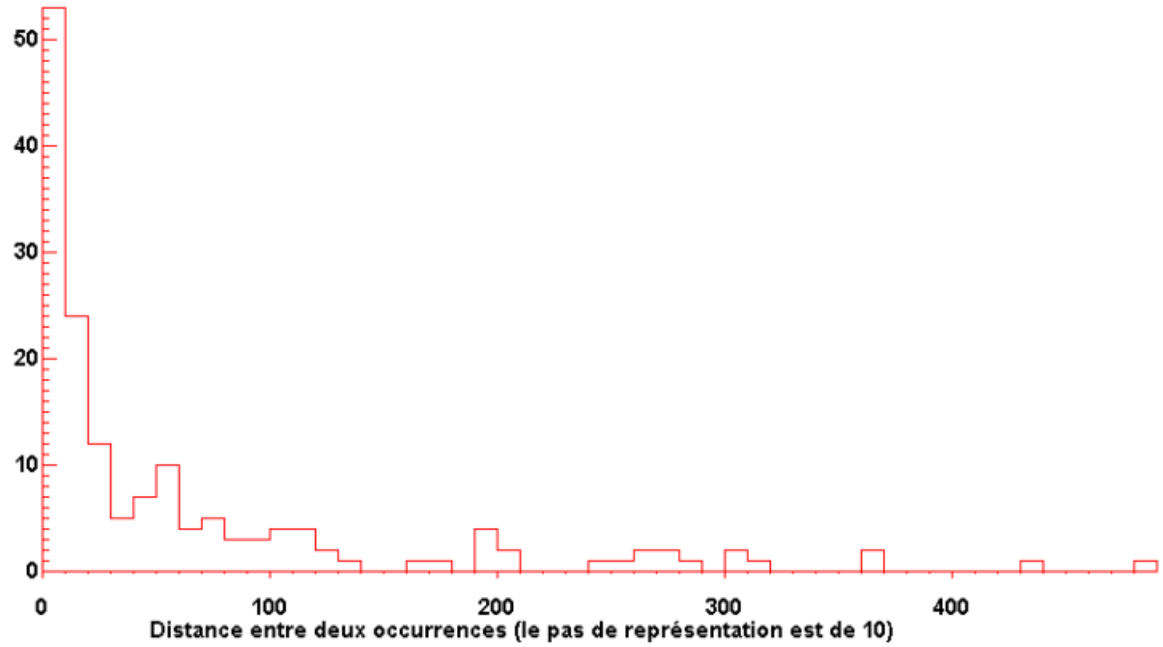


Figure 2. Graphe de répartition des imparfaits dans la Guerre d'Alexandrie

La figure 2 offre un autre mode de représentation de la même réalité, donnant à voir le nombre d'occurrences d'imparfait séparées par un intervalle inférieur ou égal à 10 formes graphiques (ici 53), le nombre de celles qui sont séparées par un intervalle compris entre 10 et 20 formes (24), etc. Les espacements réduits sont majoritaires, mais on relève aussi quelques espacements très importants (2 de plus de 370 formes, 1 de plus de 430 et 1 de plus de 480).

3.1. ***Le découpage en zones*** : la première idée est de découper chacun des textes du corpus en « zones » ou « tranches » successives, de longueur arbitraire, et de dénombrer les occurrences des différents temps verbaux des propositions principales dans chacune de ces tranches pour pouvoir ensuite traiter la chaîne numérique ainsi obtenue. Concrètement, un texte T divisé en N tranches sera caractérisé par une série de profils tels que :

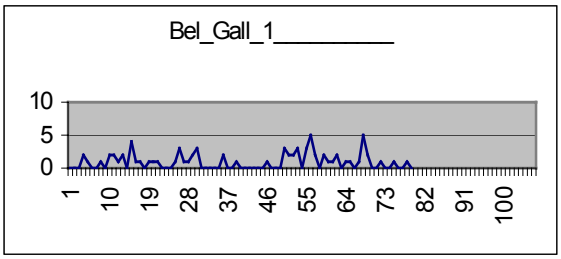
$$I_T = f_{i\,1} \quad f_{i\,2} \quad f_{i\,3} \quad \ldots \quad f_{i\,n}$$

$$P_T = f_{p\,1} \quad f_{p\,2} \quad f_{p\,3} \quad \ldots \quad f_{p\,n}$$

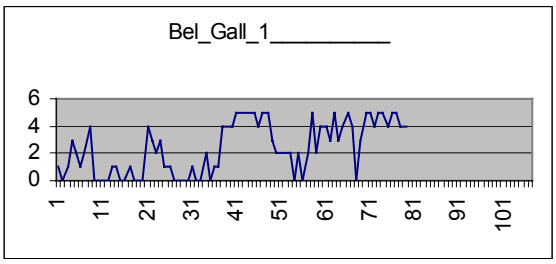
$$PR_T = f_{pr\,1} \quad f_{pr\,2} \quad f_{pr\,3} \quad \ldots \quad f_{pr\,n}$$

où $f_{i\,n}$ note la fréquence de l'imparfait dans les tranches 1, 2, 3 jusqu'à N du texte T, où $f_{p\,n}$ note la fréquence du parfait dans les tranches 1 à N du même texte et où $f_{pr\,n}$ note les fréquences du présent.

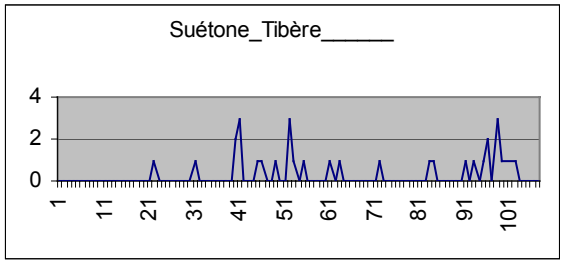
Une première représentation des données peut être réalisée sous formes de courbes grâce au simple outil Excel. Les graphes ci-dessous en donnent quelques exemples : deux textes, le livre 1 de la *Guerre des Gaules* et la *Vie de Tibère* de Suétone ont été réduits à la succession des codes caractérisant le temps et le mode de leurs verbes principaux, puis découpés en tranches de 5 codes chacune ; le premier est constitué de 79 tranches de 5 codes, le second de 107 tranches. On a décompté d'abord les occurrences du code d'imparfait, ensuite celles du code de parfait (dans chaque tranche leur nombre peut évidemment varier entre 0 et 5) :



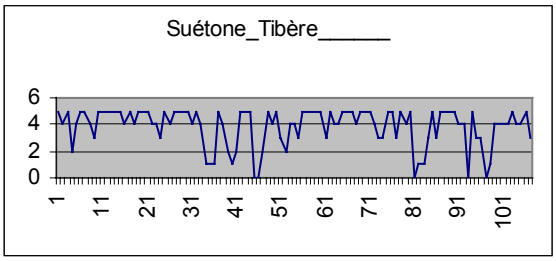
Imparfaits dans GALL_1



Parfaits dans GALL_1



Imparfaits dans TIBERE



Parfaits dans TIBERE

Cette représentation a le mérite d'être plus parlante qu'un tableau numérique. On y constate la relative complémentarité de la distribution de l'imparfait et du parfait, surtout dans le livre 1 de la *Guerre des Gaules* où la plupart des pics de l'un des temps correspondent aux creux de l'autre. On y voit que ni dans l'un, ni dans l'autre texte l'imparfait n'abonde dans les premiers chapitres, contrairement à ce que l'on pourrait attendre d'un temps descriptif d'arrière-plan destiné à poser le cadre des actions à venir. On y voit surtout la différence d'écriture entre les deux textes, avec un emploi beaucoup plus monotone du parfait dans la *Vie de Tibère* par Suétone. Ces analyses restent cependant très approximatives et de portée limitée.

On peut en outre se demander si cette méthode des tranches est tout à fait satisfaisante ; elle implique en effet une conception discontinue de la chaîne linéaire du texte. C'est pourquoi on estime pouvoir sensiblement améliorer la représentation en recourant à une mesure de densité des occurrences qui prend appui sur une véritable topologie discrète des textes.

3.2. ***La mesure de voisinage*** : il s'agit d'affecter à chaque unité constitutive du texte (c'est-à-dire, rappelons-le, à chaque code de temps verbal) une mesure de son voisinage calculée en fonction d'une propriété de ce voisinage jugée pertinente. Là encore, comme précédemment pour la longueur des tranches, la taille du voisinage peut être parfaitement arbitraire. Nous avons choisi une taille de 11 unités. Au sein de ce voisinage on a dans un premier temps retenu la propriété de présence du code 14 affecté au parfait de l'indicatif, temps fondamental de la narration. Ainsi, pour chaque unité du texte, un programme examine les cinq codes précédents et les cinq codes suivants et décompte le nombre d'occurrences du code 14 : nous donnons à cette mesure le nom de ***densité***. Le résultat aboutit à une représentation topologique discrète du texte qui rend mieux compte que la méthode précédente de sa continuité

puisque la même mesure est affectée à chaque unité successivement. Voici les résultats obtenus sur le début du livre 2 de la *Guerre Civile* :

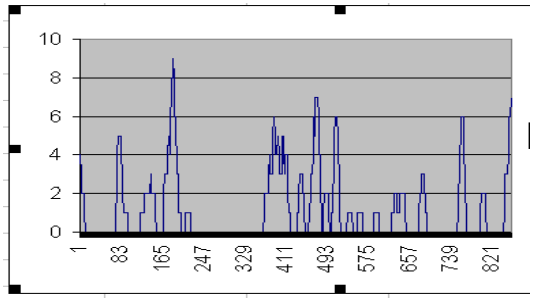
Chaîne linéaire des codes de temps et de modes affectés à chacun des verbes principaux :

11 12 11 11 11 11 11 11 12 12 12 12 12 12 12 12 12 12 11 11 11 11 11 15 15 12 15 15 14 11 11 15 11 11 11 11 11 11 11 15 15 15 11 12 12 15 14 12 12 12 12 12 12 14 14 14 14 11 11 14 14 12 14 14 14 14 14 14 14 14 14 14 12 12 12 12 14 14 14 14 14 14 14 14 14 14 14 14 15 14 14 12 12 12 12 14 14 etc.

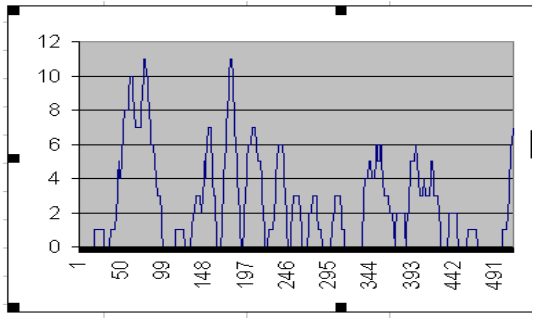
Topologie établie à partir des mesures de voisinage de chaque unité (taille du voisinage = 11, propriété du voisinage = nombre d'occurrences du code 14) :

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 2 3 4 5 4 4 5 6 6 7 8 8 8 8 8 9 10 10 10 10 9 8 7 7 7 7 7 7 7 7 8 9 10 11 11 10 10 10 9 8 7 6 6 6 6 etc.

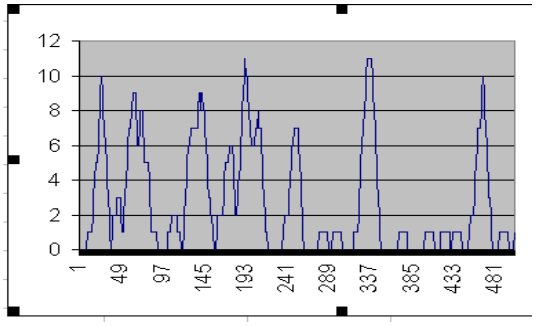
Cette nouvelle suite numérique peut de nouveau donner lieu à une représentation graphique sous forme de courbe. Voici celles de la *Guerre Civile* livres 1 et 2, la *Guerre des Gaules* livre 5, la *Guerre d'Espagne*, des livres 3 et 5 des *Histoires* de Quinte-Curce, des *Vies* de Iulius et de Tibère par Suétone et des livres 12, 13, 14 et 15 des *Annales* de Tacite.



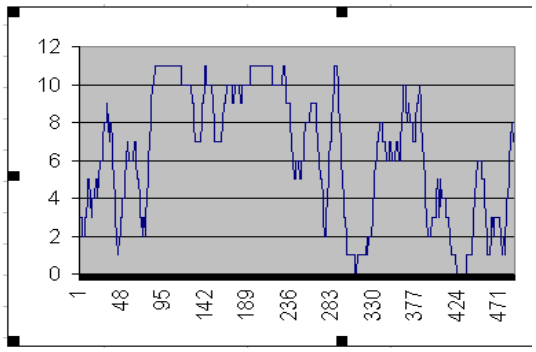
Guerre Civile 1 (César)



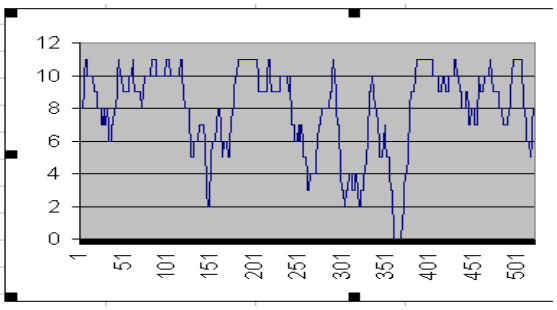
Guerre Civile 2 (César)



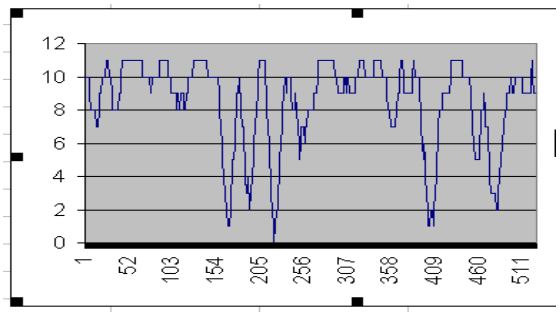
Guerre des Gaules 5 (César)



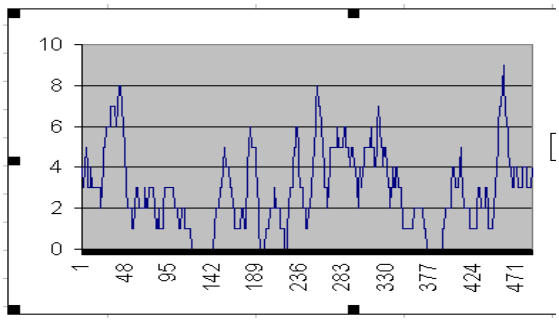
Guerre d'Espagne (imitateur de César)



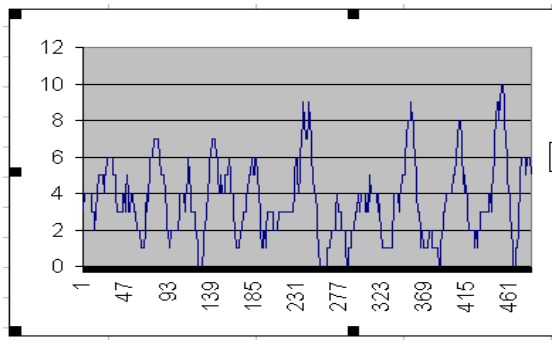
Vie de Iulius (Suétone)



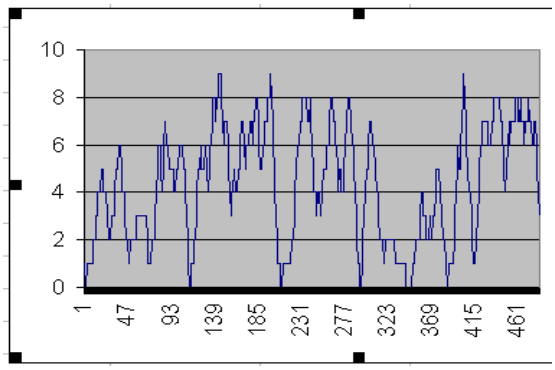
Vie de Tibère (Suétone)



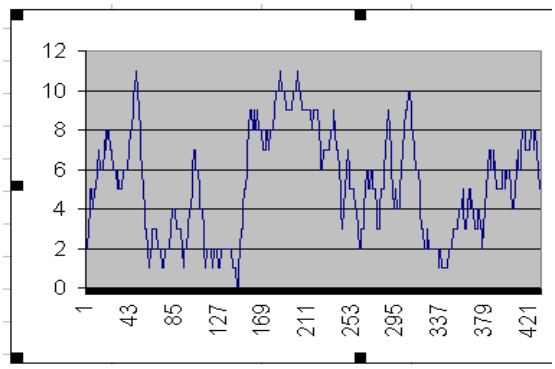
Quinte-Curce livre 3



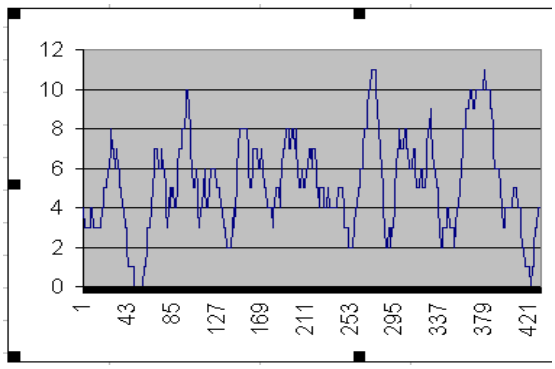
Quinte Curce livre 10



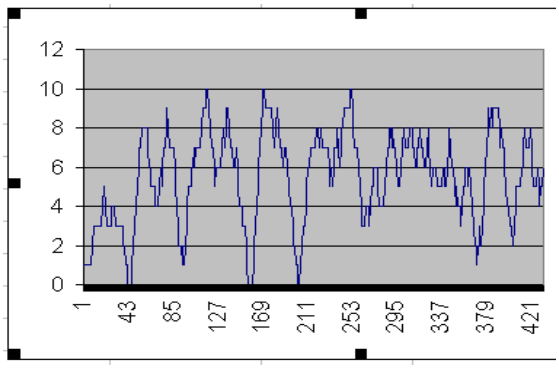
Annales 12 (Tacite)



Annales 13 (Tacite)



Annales 14 (Tacite)



Annales 15 (Tacite)

Là encore les courbes sont assez parlantes ; sans entrer dans une analyse approfondie, remarquons simplement la différence très nette entre les courbes des deux livres de César et celle de son imitateur, auteur de la *Guerre d'Espagne*. Remarquons aussi le caractère très particulier des courbes des *Vies* de Suétone. Notons que les livres de César commencent tous sur un mode qui tend à exclure le parfait alors que ceux de Suétone et, dans une moindre mesure, ceux de Quinte-Curce et de Tacite acceptent plus volontiers ce temps verbal dès les premiers paragraphes.

Cependant ces courbes restent peu exploitables en l'absence de méthode pour comparer deux courbes et surtout des courbes de longueur différentes. C'est pourquoi il nous a paru indispensable de revenir à des méthodes d'analyse moins qualitative des tableaux de données numériques.

4. Les calculs de distances et les analyses arborées

4.1. ***Le problème de la longueur*** se pose aussi, à dire vrai, pour les tableaux numériques. On touche là l'une des deux difficultés majeures auxquelles nous avons été confrontés et qui, toutes deux, concernent le découpage du texte :

– d'une part ce découpage ne peut pas être fait selon un empan fixe puisque le nombre de colonnes constitutives des différents profils de textes serait alors variable en fonction de la longueur des textes (un texte contenant 300 verbes principaux et un texte en contenant 350, découpés l'un et l'autre en tranches de 10, auraient, pour l'un, un profil de 30 tranches, pour l'autre, un profil de 35 tranches) ;

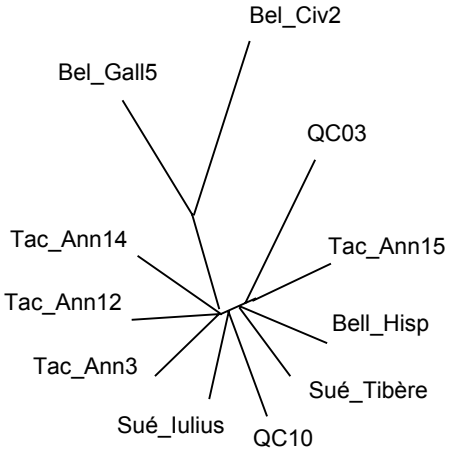
– d'autre part, dans le cadre d'un découpage devant aboutir à un nombre fixe de tranches (et donc de colonnes dans la matrice), quel est le meilleur critère de division ? On peut songer à un découpage « naturel » (du type : introduction – narration – péripétie – dénouement – conclusion). Cela pourrait être pertinent pour certains types de textes : contes traditionnels, notices explicatives ou modes d'emploi, exposés scientifiques, et même, dans le domaine latin, prose oratoire. Mais ce choix semble totalement inapplicable à textes qui sont, précisément, de structures variées, voire peu structurés. Par ailleurs, même dans l'hypothèse où un tel découpage paraîtrait praticable, on peut se demander s'il n'introduirait pas *de facto* dans les données ce qu'on y cherche, viciant ainsi dès le début toute la procédure d'analyse.

Il faut donc revenir à des tranches arbitraires et discontinues. La notion de paragraphe n'a aucun sens en latin, langue dont les scribes pratiquaient la *scriptio continua*. Celle de phrase est tout autant une projection des éditeurs modernes sur un texte en général non ponctué, mais au moins s'appuie-t-elle le plus souvent sur des critères syntaxiques et sémantiques assez fiables. Elle pourrait donc être adoptée. Cependant, étant donné la réduction que nous avons opérée sur les textes, qui les condense en une série de codes symbolisant les verbes principaux, il nous a semblé plus facile et plus pertinent de retenir cette dernière unité d'analyse comme unité de découpage (elle coïncide partiellement avec la notion de phrase, mais pas entièrement dans la mesure où une seule phrase peut être formée à partir de plusieurs propositions principales coordonnées ou juxtaposées). Nous avons également restreint le corpus à des textes de taille sensiblement égale (la longueur varie d'un facteur 1,2[2]).

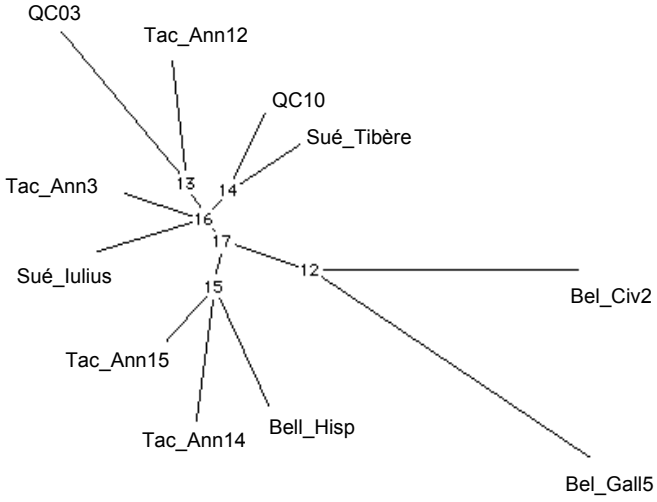
Nous avons ensuite choisi d'établir le fractionnement des textes en un nombre de tranches arbitraire qui varie de manière décroissante : les textes ont d'abord été découpés en 20 tranches, puis en 16, en 12, en 8, en 6 et enfin en 5 tranches. La réitération des calculs en faisant régulièrement et progressivement varier le nombre de tranches s'appuie sur l'espoir d'une part de trouver empiriquement la division la mieux adaptée à l'ensemble du corpus, d'autre part — et surtout, vu l'objectif à long terme de notre étude — de détecter les divisions les plus idoines pour chacun des sous-ensembles du corpus : en effet, il n'est pas exclu *a priori* que tel type de texte ou d'écriture révèle mieux son rythme narratif dans un profilage aux séquences très brèves et que tel autre s'insère mieux dans un profilage aux séquences plus longues. Le nombre d'occurrences du code 14 dans chaque tranche de chaque texte est comptabilisé, fournissant ainsi un profil du texte. Chacune des matrices de profils textuels ainsi obtenue a été soumise à un calcul de distance par le chi² et représentée graphiquement par une analyse arborée.

[2] Bel_Civ_2 : 520 codes verbaux ; Bel_Gal_5 : 512 ; Bel_Hisp : 496 ; QC03 : 558 ; QC10 : 492 ; Tac_Ann3 : 541; Tac_Ann12 : 519 ; Tac_Ann14 : 531 ; Tac_Ann15 : 575 ; Sué_Tibère : 536 ; Sué_Iulius : 529.

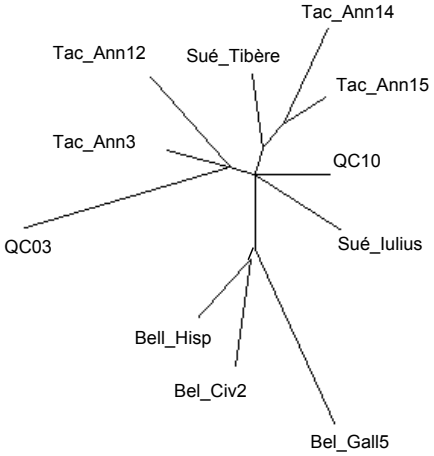
4.2. Premiers résultats et perspectives



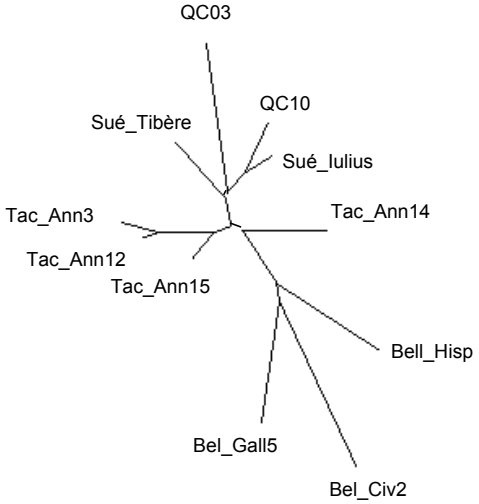
Découpage 20, code 14



Découpage 16, code 14

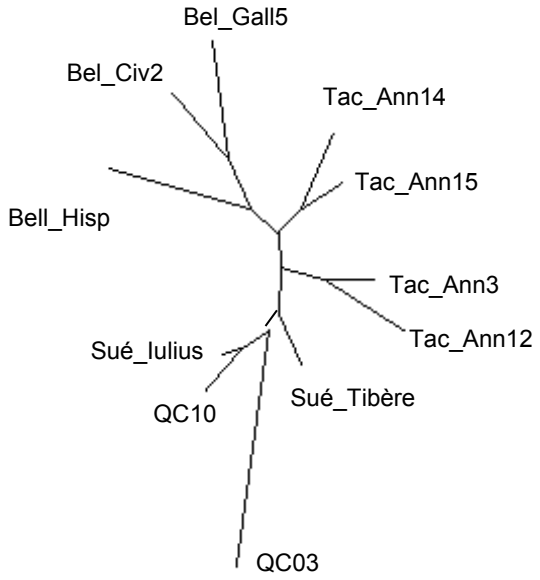


Découpage 8 , code 14



Découpage 6, Code 14

Comme on le voit sur les graphes ci-dessus et sur le graphe ci-dessous, le « découpage 20 » ne donne lieu qu'à un seul regroupement pertinent, celui des deux livres de César (qui reste stable à travers tous les découpages). Le reste prend la forme d'une étoile, dénotant l'absence de structure et de tout autre regroupement pertinent. Le « découpage 16 » ébauche une forme un peu mieux structurée, notamment autour du nœud 15, sans être toutefois très claire. Il faut en fait attendre les « découpages 6 et 5 » pour obtenir des arbres parfaitement structurés et interprétables, dans lesquels on observe un très net regroupement des œuvres par auteur.



Découpage 5, Code14

C'est donc un découpage en un nombre restreint de tranches qui permet de mettre en évidence cette structuration ; cela veut dire que le caractère dominant qui détermine la similarité ou la dissemblance entre les textes est leur structuration en larges zones correspondant à l'organisation du récit. Ce caractère semble devoir être attribué au sous-genre narratif (commentaire, biographie, etc.) plutôt qu'au style d'un auteur ; un bon indice en est la proximité récurrente des textes de Quinte-Curce et de Suétone, qui comportent les uns et les autres une forte composante biographique.

5. Conclusion

Les deux méthodes utilisées, — découpage du texte en un certain nombre de zones et mesure de voisinage —, mettent en évidence des aspects qui, tantôt rapprochent les textes, tantôt les séparent. La mesure de voisinage est un outil souple permettant de réduire la part d'arbitraire du descripteur. Cette méthode a l'avantage d'allier approche quantitative et approche qualitative dans la mesure où la qualité définitoire du voisinage peut être très variée : comme propriété pertinente, nous avons retenu ici la simple présence d'occurrences d'un temps donné au sein du voisinage, mais on peut également dénombrer, au sein du voisinage, des séquences diverses, comprenant ou excluant la présence de telle ou telle occurrence ; en outre des seuils de fréquences peuvent être fixés. La propriété du voisinage peut donc être associée à un objet de nature complexe. Toutefois, quand il s'agit de comparer des textes de longueurs différentes, la méthode présente un inconvénient majeur : le manque d'instruments de comparaison des courbes de distribution obtenues (à cet égard, toute suggestion serait la bienvenue). Le découpage du texte en un nombre fixe de tranches est une méthode moins subtile, mais cette segmentation présente l'avantage d'aboutir à un nombre fixe de colonnes dans la matrice et de permettre, pour la mesure des distances intertextuelles, d'utiliser l'outil éprouvé que constitue l'analyse arborée. En outre, cette méthode tend à mettre en évidence la structure même des textes, plutôt que les techniques d'écriture propres à chaque auteur. Une piste pour associer les avantages de chacune des deux méthodes serait de confier à un programme informatique le soin de découper la série numérique qui rend compte de la topologie du texte en N tranches respectant au mieux les variations de densité observées. Il s'agirait en quelque sorte de pro-

duire une structure en N tranches approximant au mieux la structure réelle de la topologie. Une recherche en ce sens est actuellement en cours.

Références

Évrard Ét. et Mellet S. (1998). Les méthodes quantitatives en langues anciennes. LALIES, vol. (18) : 111-155.

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des JADT 2002* : 403-411.

Longrée D. et Luong X. (à paraître, 2003). Spécificités stylistiques et distributions temporelles chez les historiens latins. In Williams G. (Ed.), *Actes des 2èmes Journées de la Linguistique de Corpus (Lorient, 12-14 sept. 2002)*. A paraître aux Presses universitaires de Rennes.

Longrée D. (à paraître, 2004). Temps verbaux et spécificités stylistiques chez les historiens latins. In Calboli G. *et al.* (Eds), *12ème Colloque intern. de Linguistique latine* (Bologne, 9-14 juin 2003).

Longrée D. et Luong X. (2003). Temps verbaux et linéarité du texte : recherches sur les distances dans un corpus de textes latins lemmatisés. *Corpus*, vol. (2) : 119-140.

Luong X. et Mellet S. (1995). Les calculs multidimensionnels au service de l'analyse syntaxique diachronique. In Bolasco S., Lebart L. et Salem A. (Eds), *Analisi statistica dei Dati testuali*. CISU : 281-288.

Luong X. et Mellet S. (2003). Mesures de distance grammaticale entre les textes. *Corpus,* vol. (2) : 141-166.

Mellet S. (2002). La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte. *Médiévales*, *Le latin dans les textes*, vol.(42) : 13-26.

Pincemin B. (2002). Similarités texte – textes. Expérience d'une application de diffusion ciblée et propositions. In *Actas del segundo Seminario de la Escuela interlatina de altos Estudios en Lingüística aplicada, Matemáticas y Tratamiento de Corpus* (s.n) : 35-52.

Salem A. (2002). Topographie textuelle dans l'analyse quantitative des textes. In *Actas del segundo Seminario de la Escuela interlatina de altos Estudios en Lingüística aplicada, Matemáticas y Tratamiento de Corpus* (s.n.) : 53-59.

L'évolution en français de l'adjectif épithète vers la post-position : réalité syntaxique ou trompe-l'œil lexical ?

Jean-Luc Manguin

CRISCO, CNRS UMR 6170 – Université de Caen, 14032 CAEN Cedex – manguin@crisco.unicaen.fr

Abstract

This study bares on the temporal variations of the position of attributive adjectives in French, in the category-tagged corpus Frantext, between 1830 and 1959, and in four different literary genders ; after we have noticed the evolution of the position of attributives, we propose four hypothesis to explain the observed variations, then we examine them. At the end, it appears that one of them (based on the variations of use) explains most of the observed effects. As a final point we propose to explore a stylistic hypothesis to go further in the explanation of these variations of use.

Résumé

Cette étude porte sur les variations chronologiques de la position de l'adjectif épithète en français, dans le corpus Frantext catégorisé, entre 1830 et 1959, et ceci dans quatre genres littéraires ; après avoir constaté l'évolution de la position de l'épithète, nous proposons quatre hypothèses pour expliquer les variations observées, puis nous les examinons. Il s'avère en conclusion que l'une d'entre elles (qui concerne les variations d'emploi) explique la majorité des effets observés. Nous proposons au final une piste à caractère stylistique pour expliquer ces variations d'emploi.

Mots-clés : adjectif épithète, variation de position, corpus français écrit.

1. Introduction

La question de la position de l'adjectif épithète en français a déjà été étudiée, engendrant une littérature abondante (voir notamment Reiner, 1968 ; Forsgren, 1978 ; Wilmet, 1981 ; la synthèse de Goes, 1998 : 77-107 ; ou encore le point de vue récent de Abeillé et Godard, 1999) ; cela dit, si la plupart des auteurs s'attachent à expliquer la position de l'épithète dans un cadre synchronique, la question de son évolution chronologique éventuelle depuis le 19e siècle jusqu'à notre époque reste en revanche relativement peu abordée[1]. Et par ailleurs, il est parfois frustrant d'entendre affirmer que « l'adjectif évolue vers la postposition » sans que rien d'autre qu'un comptage d'occurrences ne vienne véritablement expliquer cette constatation, ce qui laisse le lecteur quelque peu sur sa faim (Damourette et Pichon, 1911-1930 : § 528). Grâce à l'arrivée de corpus étiquetés, et de moyens informatiques, le traitement statistique de données textuelles permet d'étudier ce problème non seulement sur un nombre d'occurrences beaucoup plus élevé, mais encore de l'examiner du point de vue du lexique, ce qui n'avait été fait que pour les couples d'unités les plus employées ; l'apport de la lemmatisation est fondamental dans ce cas, et c'est, comme nous le verrons, ce qui permet de dégager un premier niveau de causes pour la période considérée.

[1] Il faut cependant signaler l'étude de Hug (1971) et préciser que l'étude de Glatigny (1967) intègre l'aspect temporel, mais pour un seul auteur (Nerval).

2. Étude des occurrences

2.1. *Recueil des données*

Nous nous servons du Corpus Frantext catégorisé, dans la période 1830-1959, et dans les quatre genres suivants : Mémoires, Romans, Essais & Traités, et Théâtre ; l'ensemble de ces textes compte 90 millions de mots répartis dans environ 1300 œuvres ; si l'on veut découper cet ensemble en trois tranches à peu près régulières de 30 millions de mots chacune, on aboutit alors à un partage en trois périodes : 1830-1859, 1880-1929 et 1930-1959. Cela dit, un examen préalable du corpus nous révèle deux choses :

1) Le corpus Frantext n'est pas homogène en genre tout au long de la période choisie, comme le prouve la proportion de chaque type de texte dans les trois tranches temporelles[2] :

	1830-79	**1880-1929**	**1930-59**
essais + traités	16,2 %	23,4 %	41,2 %
romans	62,0 %	55,5 %	41,2 %
mémoires	16,6 %	16,0 %	11,2 %
théâtre	5,1 %	5,1 %	6,4 %

Ce qui est ici flagrant, c'est l'énorme variation de la proportion des textes techniques (Essais & traités) au cours des trois périodes.

2) La proportion moyenne d'épithètes postposées sur toute la période d'étude varie suivant le genre de texte, comme l'avait déjà noté Larsson (1994) ; le tableau ci-dessous donne les valeurs dans notre corpus :

	proportion moyenne
essais + traités	71,6 %
romans	54,8 %
mémoires	57,5 %
théâtre	46,7 %

On se rend compte à partir de là, qu'il est donc nécessaire d'étudier séparément chaque genre littéraire, et de relever dans chacun le nombre d'épithètes anté- ou postposées ; cette étude, faite par tranches décennales[3], nous donne les résultats ci-dessous, après régression linéaire des données :

	proportion moyenne	**variation estimée**	**coefficient de détermination[4]**
essais + traités	71,6 %	10,2 %	0,55
romans	54,8 %	6,5 %	0,69
mémoires	57,5 %	6,7 %	0,39
théâtre	46,7 %	3,7 %	0,05

[2] Le calcul est effectué sur le nombre de mots.
[3] Nous donnons en annexe 1 les chiffres détaillés, excepté pour le théâtre.
[4] Il s'agit ici du coefficient r^2 de la régression linéaire.

Ce tableau synthétique nous indique que la progression vers la postposition est sensible dans tous les genres[5] ; toutefois, dans le théâtre, les données relevées dans Frantext ne permettent pas de conclure à une progression linéaire, ni même à une progression tout court. Pour les trois périodes « de base », les occurrences pour ce genre sont en effet :

théâtre	**1830-1879**	**1880-1929**	**1930-1959**
œuvres	67	64	71
mots	1477033	1553808	1957934
occurrences anté	12165	13058	15840
occurrences post	9327	13239	13471
proportion post	43,4 %	50,3 %	46,0 %

autrement dit, une évolution « en cloche » qui ne révèle pas de tendance globale.

Dans la suite de cette étude, nous nous limiterons donc aux trois genres non dramatiques.

2.2. *Quantification des erreurs*

Chaque utilisateur de la base Frantext catégorisée a pu constater que, à l'instar de ses « confrères », l'étiqueteur utilisé pour Frantext commet un certain nombre d'erreurs, que certains ont déjà étudiées en détail dans des cas précis (voir Hug, 2002). Dans notre cas, les erreurs sont, comme nous allons le voir, plutôt rares ; ceci vient du fait que nous imposons la cooccurrence de deux catégories. Ainsi, en antéposition, nous demandons au moteur de recherche de Frantext de nous trouver une structure Adjectif + Nom (ce qui donne la formulation *&e(g=A) &e(g=cS)* dans l'interface de Frantext). Seule la fonction d'épithète peut ainsi être captée, et les éventuelles erreurs ne proviennent que de confusions entre homographes.

Pour la postposition, la requête formulée Nom [+ Adverbe] + Adjectif (soit dans Frantext *&e(g=cS) &?(&e(g=Adv)) &e(g=A)*), capte non seulement la fonction d'épithète, subissant également les erreurs d'étiquetage, mais aussi l'attribut du complément d'objet (qui reste assez peu fréquent) et la fonction attributive dans le cas d'une accumulation provoquant l'ellipse de la copule (voir exemple ci-après).

Pour estimer les erreurs par la méthode d'échantillonnage, nous avons vérifié manuellement un échantillon que nous avons prélevé ainsi : pour chaque tranche décadaire, nous avons conservé les 50 premières occurrences, ce qui nous donne 650 exemples ; cet échantillonnage est effectué séparément pour chaque position adjectivale, car nous supposons, pour les raisons détaillées précédemment, que le taux d'erreurs ne sera pas le même dans les deux cas.

Voici les résultats concernant la partie « Mémoires » du corpus :

	Antéposition	**Postposition**
Erreurs de catégorie	11	13
Erreurs de fonction		10
Taux d'erreurs	1,69 %	3,54 %
Borne inférieure	1,56 %	3,28 %
Borne supérieure	1,82 %	3,80 %

[5] La limite de 1959 a été choisie en raison de l'absence de textes du genre Mémoires au-delà de cette date ; néanmoins, une étude avec une date limite ultérieure, dans les genres Roman et Essais & traités, amène une amélioration du coefficient de la régression linéaire.

Pour évaluer les bornes entre lesquelles se situe le taux d'erreur avec une probabilité de 95 %, nous utilisons la méthode classique de Student (on peut se reporter, pour le détail du calcul, à Dodge, 1999 : 257-259).

Comme on le voit, le taux d'erreur est faible, et surtout, si l'on soustrait le nombre d'erreurs estimées à nos résultats, les variations probables (0,13 % pour l'antéposition et 0,26 % pour la postposition, ce qui fait 0,4 % de précision sur la proportion) restent négligeables au regard des variations chronologiques observées.

Pour entrer dans les détails, voici quelques erreurs relevées ; tout d'abord un cas d'antéposition qui n'en est pas un :

ils se déclaraient hostiles au principe de la nation armée dans lequel ils voyaient l'amorce d'une ***armée milicienne*** *;*

(Maréchal Joffre, *Mémoires 1910-1917*, 1931, Page 8)

Ensuite, l'inverse, c'est-à-dire une antéposition réelle détectée comme postposition :

Dans ce pays de Valois, l'été n'est que le ***trouble confluent*** *du printemps et de l'automne.*

(François Mauriac, *Journal III*, 1940, Page 209)

Enfin, voici un cas d'accumulation de plusieurs structures attributives, où l'absence de la copule fait entrer dans nos résultats trois occurrences sans épithète :

ses connaissances étaient immenses, son ardeur d'étude infatigable, sa ***conversation brillante****, son caractère excellent.*

(Benjamin Constant, *Le cahier rouge*, 1830, Page 12)

Pour les Romans, ainsi que pour les Essais & Traités, le taux d'erreur est tout à fait comparable. Ayant ainsi validé les variations que nous avons observées, nous pouvons maintenant tenter de les expliquer.

2.3. *Hypothèses de travail*

Revenons à nos chiffres, qui révèlent une progression vers la postposition, dans le prolongement sans doute de celle déjà signalée par Damourette et Pichon, 1911-1930, qui voyaient en elle une « évolution de la langue » qui « paraît se continuer de nos jours »[6]. D'après ce que nous connaissons de l'évolution des langues en général, les effets de cette évolution qui peuvent se manifester dans notre étude sont d'ordre syntaxique ou d'ordre lexical.

Au point de vue syntaxique tout d'abord, il est généralement admis (voir Glatigny, 1967 ; Wilmet, 1981) que la position postposée est une caractéristique « moderne » du français, à laquelle certains adjectifs (les plus anciens, qui sont aussi les plus courants) résistent pour diverses raisons qui sont discutées par ces auteurs. On peut alors penser que les unités qui n'ont pas de position préférentielle nette vont, dans le mouvement évolutif, adopter à force d'usage un placement qui deviendra à la longue de plus en plus postposé. D'autre part, comme l'a constaté Glatigny (1967), le doublet {Adjectif, Nom} va généralement s'ordonner suivant les masses croissantes ; cela est bien sûr renforcé par l'adjonction de « satellites » à l'adjectif (adverbe ou complément, par exemple) qui vont accroître la masse du groupe adjectival. Forsgren (1978 : 159-162) avait lui aussi constaté la corrélation entre présence de l'adverbe[7] et position postposée ; il est alors possible d'imaginer que l'évolution se traduise

[6] Citation du paragraphe 528 du tome II de l'ouvrage « Des mots à la pensée ».
[7] À l'exception toutefois de *très*.

aussi par un emploi plus fréquent des adverbes.

Au point de vue lexical ensuite, nous savons d'une part que, selon ces mêmes auteurs, la création de nouveaux adjectifs (par dérivation, généralement) donne des unités adjectivales qui seront généralement postposées pour deux raisons : la dérivation suffixale leur confère une masse syllabique importante, et leur nouveauté ne les pourvoit que d'un sens précis et objectif (Glatigny, 1967 ; Reiner, 1968). D'autre part, si le lexique évolue dans son volume, la fréquence d'emploi de ses éléments varie suivant l'époque ; le fonds ancien des adjectifs, dont la caractéristique essentielle est d'être presque toujours antéposés, pourrait-il être victime du progrès et voir son usage décroître, en même temps que son sens se dilue, qu'il se banalise (Blinkenberg, 1933), et que son poids syntaxique s'affaiblit (Abeillé et Godard, 1999) ?

Ainsi, pour expliquer les variations que nous avons constatées, nous pouvons envisager quatre hypothèses, qui peuvent coexister, mais dont il importe de définir les rôles respectifs ; ces quatre possibilités reflètent chacune la variation d'un seul paramètre, tandis que les autres restent constants :

a) variation syntaxique : les adjectifs tendent globalement à être plus employés postposés ; par conséquent, si l'on mesure pour chaque adjectif la variation chronologique de la proportion d'occurrences postposées, la moyenne de toutes ces variations doit être positive (mouvement macroscopique qui n'exclut pas des exceptions individuelles).

b) élargissement lexical : le nombre d'adjectifs augmente, et les nouveaux venus s'emploient plutôt en position postposée. Pour valider cette hypothèse, il faut connaître le nombre d'adjectifs employés dans chaque période.

c) fréquence d'emploi : le lexique reste constant, mais les adjectifs préférentiellement postposés sont employés plus souvent (ou les antéposés moins souvent).

d) emploi d'adverbes : certains adverbes induisant la postposition sont plus fréquemment employés.

L'hypothèse d) est rapidement infirmée, car les occurrences impliquant des adverbes, en postposition comme en antéposition, sont en nombre quasi constant, voire en diminution ; on relève par exemple, pour les Romans :

adverbes	**1830-79**	**1880-1929**	**1930-59**
Anté	11112	7363	4479
Post	25727	25411	18247

soit, en pourcentage du total des occurrences :

adverbes	**1830-79**	**1880-1929**	**1930-59**
Anté	4,7 %	3,5 %	3,1 %
Post	10,1 %	9,4 %	9,6 %

Cette hypothèse doit donc être rejetée.

3. Étude des lemmes

La poursuite de notre étude s'est déroulée de la manière suivante : après rapatriement des occurrences fournies par Frantext, nous avons isolé les formes adjectivales, puis nous les

avons regroupées par lemmes au moyen d'un dictionnaire des formes fléchies ; au passage, cette lemmatisation a éliminé quelques formes rares, principalement des ordinaux composés. À l'issue de cette étape, nous avons pu comparer, pour chacun des trois genres, et pour chacune des trois époques, le nombre d'occurrences relevées et le nombre de lemmes impliqués dans les deux cas d'antéposition ou de postposition ; les relevés complets sont donnés dans l'annexe 2. Contrairement aux chiffres des cooccurrences, les données concernant les lemmes ne sont pas sensibles aux erreurs catégorielles ou fonctionnelles discutées précédemment ; en effet, les unités impliquées dans ces erreurs apparaissent également dans des occurrences « vraies », et n'ont donc pas à être éliminées.

Dans ces tableaux, nous avons fait figurer en bas le résultat du test du χ^2 (autrement dit la probabilité d'indépendance des deux caractères, qui sont ici la situation temporelle et la position de l'adjectif) ; nous remarquons d'abord que l'élargissement du lexique adjectival associé à la postposition n'apparaît pas dans les Mémoires, et que dans les deux autres genres, il va de pair avec un élargissement plus sensible du lexique associé à l'antéposition, au point que la proportion d'adjectifs qui sont employés postposés tend à diminuer (comme dans les Essais & traités). Cette variation est significative, si l'on regarde les résultats des test statistiques. Parallèlement à cela, l'emploi de l'antéposition diminue (la densité des adjectifs antéposés chute grosso modo de 15 %) tandis que l'emploi de la postposition varie de manière diverse : pas de variation pour les Mémoires, assez peu pour les Romans (+ 5 %) et beaucoup pour les Essais & traités (+ 26 %). Ces variations d'emploi sont elles aussi, très significatives.

On pourrait maintenant conclure avec une quasi-certitude que l'hypothèse B (lexique) n'est pas prépondérante, et que c'est l'hypothèse C (emploi) qui prime ; néanmoins, une certaine rigueur nous oblige à prendre avec précaution des statistiques qui incluent des adjectifs n'apparaissant qu'une fois tous les 20 ou 30 ans ; pour nous affranchir de ces critiques, nous avons repris les mêmes relevés en ne conservant que les adjectifs observés au moins 20 fois dans chaque tranche, dans l'une ou l'autre des positions[8] (pour les romans, ce seuil est fixé à 50, en raison de la taille de cette partie du corpus). Le nombre de lemmes est ainsi nettement réduit, et par conséquent, le nombre d'occurrences aussi (d'environ 20 %) ; nous donnons en annexe 3 les relevés correspondants. Cette fois, comme en témoigne le test du χ^2, le lexique ne varie pas (sauf dans les Essais & traités, mais nous y reviendrons) ; par contre, on note toujours une baisse de l'emploi antéposé, et même parfois une augmentation de l'emploi postposé (+ 10 % dans les Essais et traités). Ceci confirme donc l'hypothèse C, selon laquelle l'emploi antéposé décroît, tandis que l'emploi postposé reste stable ou augmente dans certains textes.

Le lexique antéposé impliqué dans les Essais et traités s'élargit dans des proportions significatives, et ce de manière assez singulière ; en effet, les adjectifs qui apparaissent comme antéposés dans la dernière période sans être utilisés dans cette position durant les deux premières, sont essentiellement postposés ; ils accèdent après 1930 à l'usage antéposé, assez rarement toutefois[9]. Ces adjectifs sont (pour les moins rares) : *primordial* (6 occurrences), *providentiel* (5), *précaire* (5), *graduel* (5), *gratuit* (4). En voici quelques exemples :

et, comme premier milieu social proprement humain, la famille avec sa vie propre, écono-

[8] En d'autres termes, cela veut dire que l'adjectif est employé au moins 60 fois en antéposition OU 60 fois en postposition.

[9] Il faut aussi signaler que certains adjectifs — comme *équivalent*, employé 5 fois en antéposition avant 1880 dans ce genre de textes — ne se rencontrent plus dans cette position par la suite. Il y a donc un double mouvement dans le lexique antéposé, qui résulte en un enrichissement ; par contre, le lexique postposé est remarquablement stable.

mique et spirituelle, et avec ses ***primordiales*** *exigences de droit naturel dont la législation civile fixe les modalités, mais qu'elle ne crée point.*

(Jacques Maritain, *Humanisme intégral*, 1936, Pages 89-90)

Aujourd'hui que nous avons dominé l'effroi nu des présocratiques, la fausse et ***précaire*** *sérénité de l' âge aristotélicien,*

(Emmanuel Mounier, *Traité du caractère*, 1946, Page 120)

le royaume franc de Jérusalem, bénéficiant de la division politico-confessionnelle entre le khalifat fâtimide du Caire et les royaumes turcs de la Syrie intérieure, favorisé en Syrie même par le ***providentiel*** *émiettement turco-arabe, jouant à son gré de l'anarchie musulmane, apparaissait comme l'arbitre de l' orient.*

(René Grousset, *L'épopée des croisades*, 1939, Pages 208-209).

Il n'y a aucune raison a priori de penser que cette ***graduelle*** *réduction de l'organique à l'inorganique doive un jour trouver sa limite. Le mécanisme a donc actuellement une position extrêmement solide,*

(Jean Rostand, *La vie et ses problèmes*, 1939, Pages 154-155)

regarde le méchant bien en face et il l'excuse, non pas bien qu'il soit méchant, mais parce qu'il est méchant notre frère, c'est-à-dire par scandaleuse, absurde et ***gratuite*** *charité. Le pardon est la totale franchise,*

(Vladimir Jankélévitch, *Le je-ne-sais-quoi et le presque-rien*, 1957, Page 171)

Cette constatation nous ramène à l'hypothèse A (syntaxe) que rien pour le moment n'est venu infirmer ; nous avons en effet constaté que les lexiques associés à l'antéposition ou à la postposition varient en effectif. Dans le cas des essais et traités, nous avons même précisé que ce sont des adjectifs essentiellement postposés qui accèdent à l'antéposition[10]. Cela laisse sous-entendre que des variations syntaxiques entrent en jeu (même si dans le cas présent, elles vont à l'encontre de la tendance qui résulte des variations d'emploi).

Pour les adjectifs courants déjà sélectionnés, nous avons alors mesuré individuellement la variation de proportion postposée ; ensuite, pour chaque genre, et sur l'ensemble des variations mesurées, nous avons calculé les paramètres statistiques élémentaires (moyenne, médiane et écart-type) ; ces calculs donnent :

	romans	**essais & traités**	**mémoires**
moyenne	1,04 %	0,99 %	-2,22 %
médiane	0,00 %	0,00 %	-0,56 %
Écart-type	8,56 %	8,61 %	10,42 %

Cela montre que globalement, la variation syntaxique est très faible sur la période considérée ; néanmoins, les variations individuelles ne sont pas inexistantes, comme le prouvent les valeurs des écarts-types. Une étude plus détaillée, que nous ne pouvons malheureusement insérer dans le cadre de cet article, montre en outre que ce sont les adjectifs dont la proportion moyenne de postposition est proche de 50 % qui varient le plus. A contrario, les adjectifs fortement postposés ou fortement antéposés ne varient quasiment pas.

[10] On peut néanmoins se demander si le nombre de textes présent dans la base Frantext pour ce genre et cette période ne joue pas un rôle.

En quelque sorte, les variations syntaxiques individuelles vont globalement se compenser, malgré la variabilité extrême de certains adjectifs, comme par exemple pour le Roman, *fatal* (qui passe de 52 à 88 % d'occurrences postposées) ou à l'inverse *bref* (de 80 à 33 %) ou encore *extraordinaire* (de 99 à 67 %).

4. Conclusion

En conclusion, et même si nous avons en fin de compte retenu essentiellement l'hypothèse C (variations d'emploi), nous pouvons dire que les variations syntaxiques recèlent peut-être l'explication au niveau suivant ; en effet, il s'avère que les diminutions d'emploi sont souvent subies par des adjectifs très courants et antéposés, comme *grand*, *jeune*, *bon*, *beau*, ou encore *vieux* ; le fait qu'un adjectif comme *extraordinaire*, emphatique de *grand*, voie dans la même période ses occurrences antéposées multipliées par 20 peut suggérer que des emplois d'adjectifs courants ont pu être transférés vers des synonymes plus précis, mais pas toujours antéposables (alors que les épithètes les plus usuelles sont très souvent antéposées notamment du fait de leur brièveté, voir Guiraud, 1960 ; Glatigny, 1967 ; Muller, 1979), ou bien des synonymes antéposables, mais avec un effet de désémantisation (voir Blinkenberg, 1933 ; Wilmet, 1981 ; Goes, 1998 : 90-97).

D'autre part, il serait intéressant d'explorer plus en détail les textes théâtraux, pour tenter d'expliquer tout d'abord pourquoi on y rencontre trois fois moins d'épithètes que dans les autres genres, et ensuite pourquoi la proportion d'occurrences postposées, qui ne semble pas augmenter, y est aussi nettement plus faible (voir, là aussi, Muller, 1979, sur la lexicométrie des textes théâtraux).

Références

Abeillé A. et Godard D. (1999). La place de l'adjectif épithète en français : le poids des mots. *Recherches linguistiques*, vol. (28) : 9-31.

Blinkenberg A. (1933). *L'Ordre des mots en français moderne*. Levin & Munksgaard.

Damourette J et Pichon E. (1911-1930). *Des mots à la pensée,* t. (II). D'Artrey.

Dodge Y. (1999). *Premiers pas en statistiques*. Springer.

Frantext, base textuelle catégorisée (1999). CNRS, ATILF (Analyse et traitement informatique de la langue française), UMR CNRS-Université Nancy2, http://www.inalf.fr/atilf .

Forsgren M. (1978). *La Place de l'adjectif épithète en français contemporain*. Studia Romanica Upsaliensia.

Glatigny M. (1967). La place des adjectifs épithètes dans deux œuvres de Nerval. *Le Français moderne,* vol. (35) : 201-220.

Goes J. (1998). *L'Adjectif entre nom et verbe*. Duculot.

Guiraud P. (1960). *Problèmes et méthodes de la statistique linguistique*. PUF.

Hug M. (1971). L'adjectif épithète et le complément du nom dans la langue des journalistes. *Études de linguistique appliquée*, vol. (I) : 58-102

Hug M. (2002). Désambiguïsation automatique d'homographes verbe/nom. In *Actes des JADT 2002*.

Larsson B. (1994). *La place et le sens des adjectifs épithètes de valorisation positive*. Lund University Press.

Muller Ch. (1979). *Langue française et linguistique quantitative*. Slatkine.

Reiner E. (1968). *La Place de l'adjectif épithète en français. Théories traditionnelles et essai de solution.* Braumüller.

Wilmet M. (1981). La place de l'épithète qualificative en français contemporain. *Revue de linguistique romane,* vol. (177-178) : 17-73.

Annexe 1. Données relevées pour les cooccurrences

Romans	**1830-39**	**1840-49**	**1850-59**	**1860-69**	**1870-79**	**1880-89**	**1890-99**	**1900-09**	**1910-19**	**1920-29**	**1930-39**	**1940-49**	**1950-59**
nb œuvres	30	93	22	22	27	46	27	30	37	94	66	44	37
nb mots	3045900	6701356	2043168	3140193	2925349	3972500	2241758	2268951	2820948	5671203	5855916	3649738	3170036
ante	42881	91518	24910	38150	36493	52046	31917	26365	34516	65812	68908	41766	34568
post	45243	92391	26640	45766	43479	66894	40472	37444	42743	83580	88851	54105	48083
total	88124	183909	51550	83916	79972	118940	72389	63809	77259	149392	157759	95871	82651
densité	0,0289	0,0274	0,0252	0,0267	0,0273	0,0299	0,0323	0,0281	0,0274	0,0263	0,0269	0,0263	0,0261
post/ante	51,3 %	50,2 %	51,7 %	54,5 %	54,4 %	56,2 %	55,9 %	58,7 %	55,3 %	55,9 %	56,3 %	56,4 %	58,2 %

Essais & traités	**1830-39**	**1840-49**	**1850-59**	**1860-69**	**1870-79**	**1880-89**	**1890-99**	**1900-09**	**1910-19**	**1920-29**	**1930-39**	**1940-49**	**1950-59**
nb œuvres	15	21	14	11	7	12	28	28	21	44	57	67	94
nb mots	887034	1457685	1066316	814161	451423	745567	1394982	1178964	902699	2949381	4041009	4116951	4500310
ante	11899	18771	14367	10850	4388	10450	16706	13745	10432	33900	48369	44007	52392
post	19135	33354	30865	21242	13904	19983	43247	36863	27160	79876	125037	134943	145073
total	31034	52125	45232	32092	18292	30433	59953	50608	37592	113776	173406	178950	197465
densité	0,0350	0,0358	0,0424	0,0394	0,0405	0,0408	0,0430	0,0429	0,0416	0,0386	0,0429	0,0435	0,0439
post/ante	61,7 %	64,0 %	68,2 %	66,2 %	76,0 %	65,7 %	72,1 %	72,8 %	72,2 %	70,2 %	72,1 %	75,4 %	73,5 %

Mémoires	**1830-39**	**1840-49**	**1850-59**	**1860-69**	**1870-79**	**1880-89**	**1890-99**	**1900-09**	**1910-19**	**1920-29**	**1930-39**	**1940-49**	**1950-59**
nb œuvres	7	8	4	9	2	1	10	11	6	16	12	8	8
nb mots	419677	1569687	665681	1548194	587676	93691	1549029	696190	895928	1676890	1581605	786357	1074318
ante	5604	21360	9500	19560	7982	1470	19746	9405	10382	17890	19805	8800	11408
post	6484	24924	11236	27412	10002	1993	24383	12218	12765	30521	25836	11201	21277
total	12088	46284	20736	46972	17984	3463	44129	21623	23147	48411	45641	20001	32685
densité	0,0288	0,0295	0,0312	0,0303	0,0306	0,0370	0,0285	0,0311	0,0258	0,0289	0,0289	0,0254	0,0304
post/ante	53,6 %	53,9 %	54,2 %	58,4 %	55,6 %	57,6 %	55,3 %	56,5 %	55,1 %	63,0 %	56,6 %	56,0 %	65,1 %

Annexe 2. Cooccurrences et lemmes employés

Romans

occurrences	1830-79	1880-1929	1930-59		lemmes	1830-79	1880-1929	1930-59
nb mots	17855966	16975360	12675690		nb mots	17855966	16975360	12675690
anté	226941	200733	138057		anté	1887	2375	2150
post	246570	263134	184757		post	4948	5643	5571
densité a (p. mille)	12,71	11,82	10,89		proportion P	72,4 %	70,4 %	72,2 %
densité p (p. mille)	13,81	15,5	14,58					
proportion P	52,1 %	56,7 %	57,2 %					
test	**0**				**test**	**0,01**		

Essais & traités

occurrences	1830-79	1880-1929	1930-59		lemmes	1830-79	1880-1929	1930-59
nb mots	4676619	7171593	12658270		nb mots	4676619	7171593	12658270
anté	54849	76057	130141		anté	1156	1705	2144
post	113810	198527	388553		post	3954	5465	6714
densité a (p. mille)	11,73	10,61	10,28		proportion P	77,4 %	76,2 %	75,8 %
densité p (p. mille)	24,34	27,68	30,7					
proportion P	67,5 %	72,3 %	74,9 %					
test	**0**				**test**	**0,103**		

Mémoires

occurrences	1830-79	1880-1929	1930-59		lemmes	1830-79	1880-1929	1930-59
nb mots	4790915	4911728	3442280		nb mots	4790915	4911728	3442280
anté	60990	54965	36741		anté	1309	1588	1301
post	77348	78356	55828		post	3953	4304	3679
densité a (p. mille)	12,73	11,19	10,67		proportion P	75,1 %	73,0 %	73,9 %
densité p (p. mille)	16,14	15,95	16,22					
proportion P	55,9 %	58,8 %	60,3 %					
test	**0**				**test**	**0,044**		

Annexe 3. Cooccurrences et lemmes employés (lemmes les plus courants)

Romans

occurrences	1830-79	1880-1929	1930-59		lemmes	1830-79	1880-1929	1930-59
nb mots	17855966	16975360	12675690		nb mots	17855966	16975360	12675690
anté	213840	188452	130430		anté	555	569	569
post	184350	192294	129367		post	622	620	620
densité a (p. mille)	11,98	11,10	10,29		proportion P	52,8 %	52,1 %	52,1 %
densité p (p. mille)	10,32	11,33	10,21					
proportion P	46,3 %	50,5 %	49,8 %					
test	**0**				**test**	**0,925**		

Essais & traités

occurrences	1830-79	1880-1929	1930-59		lemmes	1830-79	1880-1929	1930-59
nb mots	4676619	7171593	12658270		nb mots	4676619	7171593	12658270
anté	52151	71056	122937		anté	419	584	669
post	95268	148898	280392		post	804	806	808
densité a (p. mille)	11,15	9,91	9,71		proportion P	65,7 %	58,0 %	54,7 %
densité p (p. mille)	20,37	20,76	22,15					
proportion P	64,6 %	67,7 %	69,5 %					
test	**0**				**test**	**0**		

Mémoires

occurrences	1830-79	1880-1929	1930-59		lemmes	1830-79	1880-1929	1930-59
nb mots	4790915	4911728	3442280		nb mots	4790915	4911728	3442280
anté	56318	49448	33092		anté	345	372	351
post	52003	50740	37959		post	422	421	421
densité a (p. mille)	11,76	10,07	9,61		proportion P	55,0 %	53,1 %	54,5 %
densité p (p. mille)	10,85	10,33	11,03					
proportion P	48,0 %	50,6 %	53,4 %					
test	**0**				**test**	**0,728**		

La BDTS-concordances : un outil d'enrichissement de la pratique lexicographique

Chantal-Édith Masson, Hélène Cajolet-Laganière et Pierre Martel

Université de Sherbrooke – Sherbrooke –Québec – Canada
chantal-edith.masson@usherbrooke.ca, helene.cajolet.laganiere@usherbrooke.ca

Abstract

BDTS-concordances: an important technological tool for the enhancement of the lexicographic practice. The present paper lies within the research framework of the Centre d'Analyse et de Traitement Informatique du Français Québécois (CATIFQ) and, more particularly, of the FRANQUS research group from the Université de Sherbrooke. The objective of the FRANQUS project is to develop an original nomenclature of the French language used in Quebec, mainly of its standard usage. The current lecture is explicitly related to the initial exploitation of the *Banque de données textuelles de Sherbrooke (BDTS)* (The Sherbrooke textual database) using a tool that was developed and that is known as BDTS-concordances. First, we will introduce the BDTS, and then the BDTS-concordances from a methodological perspective. Second, using examples, we will illustrate how BDTS high-frequency words are treated in order to write lexicographic articles. It appears that contexts, for terms with a frequency rate inferior to 100, can use an "ordinary" treatment, meaning that the memory retention of the information was possible. On the other hand, close to 20% of the BDTS terms show a frequency rate of 100 or more. Those consequently received a uniform treatment based on an original model. This constancy is achieved through a Definition of Type of Document (DTD-XML). The data is entered directly in a computerized form which is displayed on the DTD in four main sections: 1-entry, 2-co-occurences and constructions, 3-comment, and 4-writing/prog. We will present a detailed description of this file and illustrate, using examples of high-frequency words, how the BDTS-concordances works in its present state, its contributions and limits. We will also point out the possible avenues considered for its optimization.

Résumé

La présente communication s'inscrit dans le cadre des travaux du Centre d'analyse et de traitement informatique du français québécois (CATIFQ), et plus particulièrement du groupe de recherche FRANQUS de l'Université de Sherbrooke. Le projet FRANQUS vise à l'élaboration d'une nomenclature originale du français en usage au Québec, essentiellement de son usage standard. L'objet du présent exposé concerne explicitement une première exploitation de la Banque de données textuelles de Sherbrooke (BDTS) grâce à l'outil développé communément nommé BDTS-concordances. Nous présenterons dans un premier temps la BDTS, puis la BDTS-concordances d'un point de vue méthodologique. Dans un deuxième temps, nous illustrerons, à l'aide d'exemples, le traitement des mots de fréquence élevée de la BDTS aux fins de la rédaction des articles lexicographiques. Il est apparu que les contextes, pour les vocables ayant une fréquence inférieure à 100, se satisfaisaient bien d'un traitement « ordinaire », c'est-à-dire que la rétention en mémoire des informations était possible. Par ailleurs, près de 20 % des vocables de la BDTS présentent une fréquence de 100 ou plus et ont fait ainsi l'objet d'un traitement uniforme selon un modèle original. C'est une Définition de Type de Document (DTD-XML) qui assure cette constance. Les informations sont saisies directement dans la fiche informatisée qui se déploie sur la DTD en quatre grandes sections : 1- entrée, 2 –cooccurrents et constructions, 3- remarque et 4- rédaction. Nous présenterons une description détaillée de ce fichier et nous illustrerons, grâce à quelques exemples de mots fréquents, le fonctionnement, les apports, de même que les limites de la BDTS-concordances telle qu'elle fonctionne actuellement. Nous ferons également état des pistes envisagées pour son optimisation.

Mots-clés : corpus, base de données textuelles, analyse, contextes, concordances, redondance, structuration, DTD, XML, informatisation.

1. Introduction

La présente communication s'inscrit dans le cadre des travaux du Centre d'analyse et de traitement informatique du français québécois, et plus particulièrement du groupe de recherche FRANQUS (français québécois : usage standard) de l'Université de Sherbrooke, responsable du projet « Nomenclatures, description et application dans les technologies de l'information et de la communication ». Ce projet est échelonné sur cinq ans; le groupe entreprend sa troisième année de travail. Le calendrier des activités prévoyait, au cours des deux premières années, la conception de tous les outils linguistiques (notamment, la constitution des bases documentaires) et informatiques (notamment, la construction de la plateforme informatique et des modèles de traitement) visant à l'élaboration d'une nomenclature originale du français en usage au Québec, essentiellement de son usage standard, et au support de l'activité de rédaction des articles lexicographiques. L'objet du présent exposé concerne explicitement une première exploitation de la Banque de données textuelles de Sherbrooke (BDTS) grâce à l'outil développé communément nommé BDTS-concordances. L'exploitation de banques de données textuelles est devenue un incontournable en linguistique, notamment en lexicographie.

De fait, la plupart des banques de données textuelles, tant anglaises que françaises, ont été constituées dans le cadre de projets visant à l'élaboration de dictionnaires. Néanmoins, un problème se pose : comment traiter ces masses de données textuelles, particulièrement pour les mots de fréquence élevée, de manière à faciliter le travail des rédacteurs et rédactrices et à respecter un échéancier dans un projet dictionnairique. Nous présenterons, dans un premier temps la BDTS, puis la BDTS-concordances d'un point de vue méthodologique, ce dernier point, appuyé de quelques exemples de traitement de mots de fréquence élevée de la BDTS aux fins de la rédaction des articles lexicographiques, en prenant soin de faire état des apports et des limites de l'ensemble ainsi que des projections de développement.

2. Banque de données textuelles de Sherbrooke (BDTS)

La BDTS est un corpus de textes mis sur pied par une équipe de chercheurs du Centre d'analyse et de traitement informatique du français québécois (CATIFQ). Il s'agit d'une banque de textes représentatifs des différents usages du français en usage au Québec[1], de diverses situations de communication et de divers registres de langue; elle contient plus de 37 millions d'occurrences, ceci, sans compter les corpus complémentaires[2] requis par son caractère volontairement évolutif. La BDTS est conçue de manière à réunir le plus grand nombre possible de thèmes, de discours et de niveaux de langue du français contemporain oral et écrit en usage au Québec. La diversité des textes sélectionnés vise à une « représentativité » du français en usage au Québec dans un contexte nord-américain. Il importe de préciser toutefois qu'il ne s'agit pas d'une représentativité au sens statistique du terme (échantillonnage non probabiliste raisonné par quota) : en effet, la notion de représentativité de la BDTS doit plutôt être associée à la variété des textes qui la composent et qui reflètent la langue générale (orale et écrite) de même que la langue littéraire, journalistique et relativement plus spécialisée utilisée au Québec dans différentes situations de communication. Il est entendu que tous les

[1] Pour la notion de français en usage au Québec, nous faisons référence aux textes stockés aux fins de diverses analyses et rédigés dans l'espace géographique du Québec. Ces textes sont considérés dans leur totalité.

[2] De fait, la BDTS mise à jour totalise aujourd'hui quelque 45 M de mots, ce qui augmente d'autant la tâche de traitement et justifie la mise en œuvre de moyens destinés à en réduire l'impact.

emplois lexicaux ou sémantiques rencontrés dans les textes de la BDTS n'appartiennent pas nécessairement à la langue dite « standard »[3].

2.1. *Typologie et composition de la BDTS*

Les 37 millions d'occurrences de la BDTS sont répartis selon cinq types de discours : 59 % de textes spécialisés (près de 22 millions d'occurrences), 16 % de textes littéraires (près de six millions d'occurrences), 14 % de textes journalistiques (plus de cinq millions d'occurrences), 6 % de textes didactiques (plus de deux millions d'occurrences) et 5 % de textes oraux (près de deux millions d'occurrences)[4]. La ventilation des pourcentages associés aux diverses catégories de textes correspond à l'objectif visé : avoir accès aux différents registres de langue, notamment le niveau standard, actuellement en usage au Québec. Les 37 millions d'occurrences de la BDTS sont répartis dans plus de 10 000 textes contemporains, couvrant les années 1960 jusqu'à 2002.

Le corpus de base contient environ 206 000 formes différentes ou quelque 192 000 formes si l'on exclut les chiffres. Comme le but du projet de description lexicographique concerne le français contemporain en usage au Québec, la plupart des textes sont postérieurs à 1960 (date choisie de façon arbitraire, mais correspondant au début de la Révolution tranquille et marquant le début du Québec moderne). Cependant, certains textes littéraires antérieurs à 1960 ont été sélectionnés en raison de leur valeur et de leurs qualités intrinsèques reconnues par les Québécois. La répartition des vocables selon les types de source s'énonce ainsi :

Le sous-ensemble de langue spécialisée contient des textes qui fournissent des mots de langue générale et le vocabulaire de base de nombreux domaines spécialisés. Il renferme en outre des textes nombreux et diversifiés : monographies, mémoires, thèses, rapports, livres, documents administratifs, etc. **Le sous-ensemble de langue littéraire** comprend des textes tirés de romans, d'essais, de poèmes, de nouvelles, de récits, de contes, de pièces de théâtre, etc. **Le sous-ensemble de langue journalistique** contient des articles tirés de différentes publications québécoises, comme des quotidiens (*La Presse, Le Devoir, Le Soleil, Le Droit,* etc.), des magazines ou périodiques spécialisés (*L'Actualité, Voir, Québec Science, Franc-Vert, Interface*, etc.). **Le sous-ensemble de langue didactique** comprend des textes tirés de manuels scolaires (de niveau secondaire, collégial, universitaire), de logiciels informatiques (textes de présentation et guides de l'utilisateur), de manuels pour une formation technique, etc. Enfin, **le sous-ensemble de langue orale** contient, d'une part, des échantillons de langue parlée spontanée, c'est-à-dire des transcriptions d'enquêtes sociolinguistiques orales réalisées dans différentes régions du Québec, et d'autre part, des échantillons de langue moins spontanée (contes, monologues, téléromans, textes radiophoniques et télévisés, tribunes téléphoniques, témoignages, etc.).

2.2. *Brève caractérisation de la BDTS*

La BDTS, si elle est une banque de taille relativement modeste en comparaison de celles des grandes banques européennes construites essentiellement à des fins lexicographiques (par

[3] Le mot *standard*, employé dans le contexte de ce projet de description lexicographique, désigne une catégorie de textes dont le niveau de langue est généralement soigné ou soutenu (textes littéraires reconnus, documents administratifs et législatifs soignés, manuels scolaires primés, productions scientifiques, livres et essais de tous genres publiés par des intellectuels, textes journalistiques, etc.); il s'oppose principalement à des textes oraux de niveau « familier » ou « très familier », etc.

[4] Nous tenons à remercier Marie-Claude Lavallée, Geneviève Labrecque, Chantal Fontaine, Lucie Lahaie et Katherine Pérusse, chercheures et rédactrices au groupe FRANQUS pour toutes les données et les informations qu'elles nous ont fournies.

exemple le *Trésor de la langue française*), présente un intérêt particulier en vertu de sa composition lorsqu'on la compare aux autres banques de données textuelles de la francophonie : Frantext est composée de 80 % de textes littéraires et 20 % de textes techniques, Beltext, de 100 % de textes littéraires et oraux, Québétext, de 100 % de textes littéraires et Suistext, de 100 % de textes littéraires.

Elle se distingue également à deux autres titres. En effet, outre son caractère évolutif évoqué plus haut, la BDTS est entièrement informatisée et indexée, ce qui en permet une consultation efficace et l'obtention rapide de toutes les références, les concordances et les attestations des mots. Ce n'est pas son informatisation et son indexation proprement dites qui la distinguent, mais plutôt les choix qui ont été faits au moment de son retraitement. En effet, entamée il y a plusieurs années, sa constitution s'était faite dans une coquille « propriétaire » opérable exclusivement en mode texte et selon des fonctionnalités prédéfinies. Si la manipulation des données, dans cette coquille, était économique et efficace, il demeure que cette manipulation était réservée aux initiés et que les résultats obtenus se montraient très peu « collaboratifs » (avec d'autres applications), autant de limitations difficilement acceptables. Pour ces raisons, et compte tenu du fait que tous les chercheurs et les rédacteurs de FRANQUS ne sont pas des informaticiens, mais doivent cependant interroger quotidiennement la BDTS, compte tenu aussi de notre objectif de développement d'une plateforme intégrée de traitement et d'une interface unifiée d'interrogation (toutes deux actuellement fonctionnelles et en usage), il nous fallait adopter un langage à la fois capable de soutenir une structuration (au besoin) très fine des informations, leur hiérarchisation, tout en étant non propriétaire (ISO) et convivial. Il y a 6 ans déjà, nous faisions le choix du langage de balisage structurel extensible XML dans cet esprit. Toutes nos bases, sans exception, sont structurées ainsi, les préexistantes, telles que la BDTS ayant fait l'objet d'une structuration et d'une conversion parallèlement à la poursuite de son développement. La BDTS XMLisée, dont les sorties sont d'autant plus lisibles qu'elles sont mises en forme à l'aide de feuilles de style XSL, se prête à tous les besoins puisqu'il nous est maintenant possible de développer « à la carte » toutes les routines informatiques nécessaires (en langage JAVA), le tout, sur nos propres serveurs hautement sécurisés.

3. La BDTS-concordances : une première exploitation de la BDTS

Malgré la convivialité supérieure du XML, la difficulté cognitive et fonctionnelle de la manipulation/consultation du corpus dans l'identification, la recension et l'analyse des contextes pour les mots de fréquence élevée est rapidement apparue, qu'ils s'agisse d'une sortie imprimée ou d'une sortie sur écran. Il fallait exercer un contrôle sur la masse documentaire, laquelle était imposante à cause de la redondance des informations. La BDTS-concordances constitue une première exploitation de la BDTS. Elle fournit une première analyse des contextes pour chaque mot destiné à faire partie de la nomenclature du dictionnaire en cours d'élaboration à Sherbrooke. Elle a pour but de réduire la tâche d'examen des contextes de deux manières travaillant en conjonction, soit une structuration de leur présentation de manière à faciliter l'identification des constructions et des cooccurrents les plus récurrents ainsi qu'un contrôle (au moins partiel) de la redondance des contextes pour les mots de fréquence élevée.

La première étape de ce processus a consisté en des balayages successifs de la BDTS à l'aide de la liste des vocables déjà identifiés comme devant faire partie de notre nomenclature. Cette procédure, réalisée par nos informaticiens, était destinée à alléger la tâche des rédacteurs en leur fournissant, sous une commode forme textuelle, toutes les occurrences d'un vocable, mises en évidence et « encadrées » de leurs contextes. Parallèlement, elle permettait de vali-

der les informations contenues dans les dictionnaires français et québécois (et compilées dans la Base Thésaurus documentaire) et de repérer les nouveautés, c'est-à-dire les informations spécifiques de la BDTS. L'ensemble se présentait ainsi sous la forme d'une liste non structurée :

```
un songe éveillé. Comment ne pas se |*griser d'illusions, comment ne pas renaît
i des compensations l'incitait à se |*griser des colères qu'il pouvait faire à
rait! Jamais plus il ne pourrait se |*griser de_l espérance d'un bonheur venant
chaleur de_la maison acheva de les |*griser en battant aux tempes et contre le
it du fond du seau finissait par me |*griser. Puis le lait s'ajoutant au lait,
eine. L'inconnu ne pouvait que nous |*griser. C'étaient des images d'un audacie
à l'autre, à l'ivre enfant Qui vous |*grisa de mots, certain soir triomphant. C
geaient, qui la soutenaient, qui la |*grisaient du fond de_la glace. Plus tard,
quelques gorgées de vin mousseux la |*grisaient plus_qu un demi#litre de whisky
|P73 saveur aux aliments. Ce qui la |*grisait surtout, c'était dans une glace p
en pointe, l'amiral, timidement, se |*grisait tous les soirs. Les deux |P140 so
et au_fur_et_à_mesure que le vin me |*grisait. Cela devenait grossier. En elle
is Chateaubriand», et cette idée me |*grisait. J'avais le sentiment de communie
 Il pourrait lui parler. Déjà il se |*grisait d'images et d'attouchements. Mais
ui peut se révéler particulièrement |*grisant pour ceux qui apprennent à écrire
en va et, en même temps, c$est très |*grisant parce qu'on est les premiers à le
mptueux que le Forum, ait été moins |*grisant, c$est compréhensible. Que Joe Ca
banques ont mieux résisté au climat |*grisant des marchés financiers d'alors. M
```

Figure 1. Exemple du traitement du vocable « griser » et de ses formes à la première étape

Après des tests auprès de nos rédacteurs, il est apparu que les contextes, pour les vocables ayant une fréquence égale ou inférieure à 100, se satisfaisaient bien d'un tel traitement, c'est-à-dire que la rétention des informations en mémoire, un élagage en temp réel de la redondance de même que l'identification des informations d'intérêt étaient possibles. Ce constat a conduit à la mise de côté temporaire d'un traitement plus poussé de tels mots, ce qui permettait de concentrer l'attention sur les mots de fréquences moyenne et élevée, nettement plus problématiques

Avec près de 20 % des mots de la nomenclature présentant une fréquence de 100 ou plus, le modèle de structuration à retenir pour la construction proprement dite des documents XML composant la BDTS-concordances devait permettre un traitement uniforme et conforme à nos besoins de même qu'au type de données. C'est une Définition de Type de Document (DTD-XML) qui assure cette constance et permet la construction de documents non seulement bien formés, mais valides.

Les contraintes logistiques étaient nombreuses puisque l'ensemble des bases regroupées dans la plateforme intégrée (BDI) comptent plusieurs dizaines de milliers de documents XML, dont plusieurs centaines peuvent être en circulation simutanément. D'un point de vue informatique, pour assurer la complétude du traitement de la liste de vocables composant notre nomenclature, pour connaître l'état d'avancement de chaque document (et pour éviter des erreurs telles que l'écrasement de fichiers), un fichier « squelette », préstructuré et comportant déjà certaines informations était prégénéré pour chacun des fichiers de cette liste. On remarquera que ces fichiers XML se sont vu attribuer une extension .CNC permettant de les différencier des fichiers « éponymes » des autres bases, d'où leur surnom de *fiche CNC*. Les autres informations étaient saisies directement dans la fiche informatisée qui se déploie sur la DTD. Celle-ci structure et regroupe les informations sous 4 grandes sections : (1) ***Entrée***, (2) ***Cooccurrents et constructions***, (3) ***Remarque*** et (4) ***Rédaction***. Elles se présentent ainsi :

« Entrée »

Elle se subdivise en trois sous-sections (dans les deux premières, les informations sont générées automatiquement). Il s'agit respectivement, plus précisément :

- du mot en entrée, de la catégorie grammaticale du vocable (appelé ici mot-pivot) dans le Thésaurus documentaire (TD) et de sa fréquence dans la BDTS. Il convient de préciser que le Thésaurus documentaire comprend une compilation des réseaux sémantiques des principaux dictionnaires usuels du français;
- des formes présentes en contextes. Dans cette partie se trouvent les formes du mot-pivot présentes dans les concordances, avec leur fréquence;
- des expressions, exemples et locutions (appelés « syntagmes » aux fins de notre recherche). Sous cette section sont groupés ces syntagmes rattachés au mot-pivot et répertoriés dans le Thésaurus documentaire (donc dans les dictionnaires), et présentés de façon hiérarchisée selon leurs sens et sous-sens, etc.

« Cooccurrents et constructions »

Cette partie de la fiche est celle où le rédacteur intervient le plus. Elle comporte un tableau qui rend compte des cooccurrents du mot-pivot, donnant, pour chaque cooccurrent, sa fréquence, son lemme et ses formes, extraits, cette fois, de la BDTS.

« Remarque »

Au bas de la fiche, un espace (de taille indéfinie) est prévu pour accueillir les commentaires des rédacteurs de fiches de la BDTS-concordances.

« Rédaction »

Tout comme dans le Thésaurus documentaire et les autres fichiers de la Base de données intégrée (BDI), la fiche comporte une partie « prog » dans laquelle est encapsulée l'information nécessaire au fonctionnement de la BDI.

4. Un exemple de traitement : le verbe *atterrir*

4.1. *Les modalités de la construction de la fiche CNC (BDTS-concordances)*

À une première étape, il s'agit de récupérer tous les sens déjà encodés dans la fiche XML correspondante stockée dans la Base Thésaurus Documentaire (TD) lorsqu'une telle fiche existe. En effet, tous les mots retenus à la nomenclature du dictionnaire propriétaire ne disposent pas d'une telle fiche puisque celle-ci, se rappelle-t-on, est une compilation des réseaux sémantiques établis dans un ensemble de dictionnaires existants; par conséquent, les mots spécifiques à la BDTS ne s'y trouvent pas représentés. Il s'agit d'abord de coller, dans les cases « cooccurrents » de la fiche CNC, les sens tirés du TD, afin de créer (à des fins de structuration et de hiérarchisation) les rubriques de la fiche. En une seconde étape, il s'agit, dans la case « constructions » associée à chaque rubrique, de coller un nombre de contextes (tirés de la BDTS) suffisant pour avoir une bonne idée de l'usage du mot dans ce sens.

4.2. *Le résultat du traitement et la fiche CNC ainsi constituée*

Une fois constituée, la fiche, qui fait état de la fréquence totale du mot-pivot (156), de la distribution de ses formes, du nombre et de la distribution de ses cooccurrents (259), qui ordonne également les « syntagmes » selon les sens et les sous-sens, etc., se présente ainsi lorsque automatiquement mise en forme à l'aide de la feuille de style (XSL) dans l'interface unifiée :

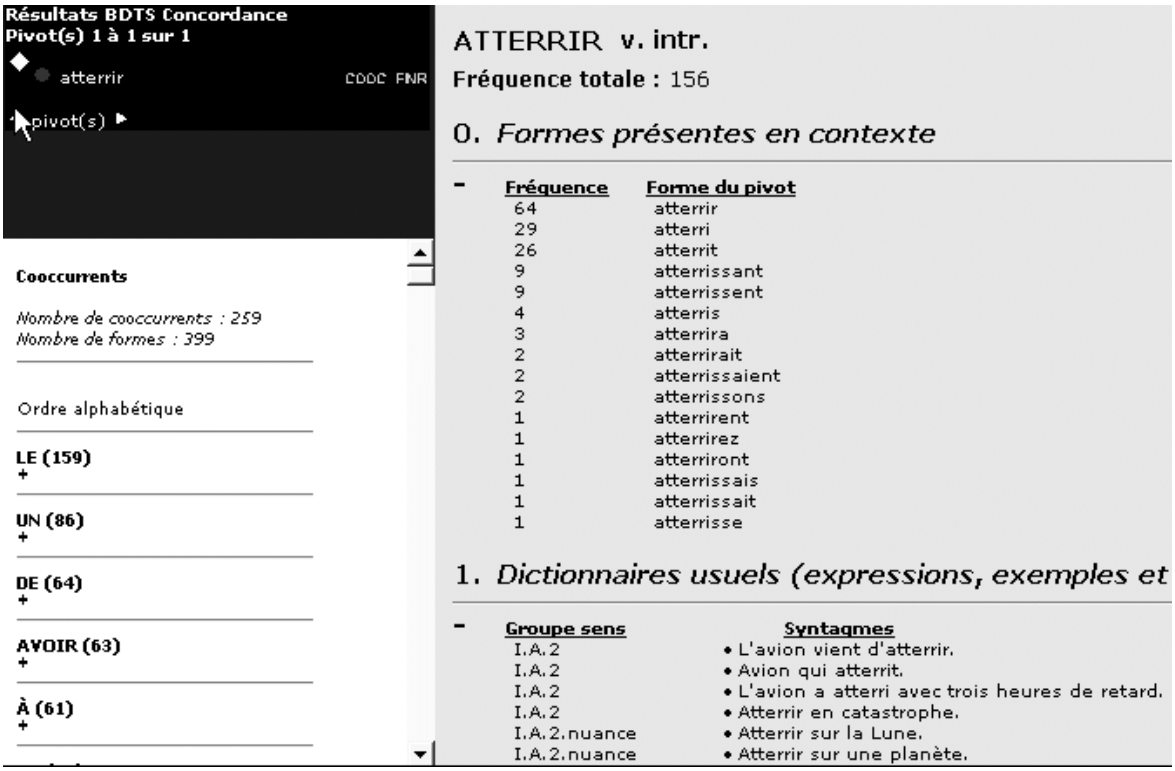


Figure 2. Vue très fragmentaire de la fiche CNC (BDTS-concordances) du verbe « atterrir »

5. Intérêt et limites de la BDTS-concordances

La construction des premières centaines de fiches CNC a été l'occasion de tester le modèle, la méthode, les instruments de même que le protocole de réalisation de la tâche et de vérifier l'adéquation et l'intérêt de cet ensemble. Les conclusions de ce processus pourraient se moduler comme suit.

5.1. *L'intérêt des fiches CNC (BDTS-concordances)*

En plus de sa très grande lisibilité et de sa richesse informationnelle, le document (la fiche informatisée complétée) qui en résulte présente l'avantage de pouvoir être intégré dans la sous-base de données (BDTS-concordances) qui lui est consacrée dans la super Base intégrée (BDI). Cette intégration en permet l'interrogation dans une interface unique, concurremment d'ailleurs, à celle de documents d'autres sous-bases également intégrées. Un tel document fournit alors rapidement, grâce à la liste structurée de cooccurrents intégrés à la fiche, des pistes pour les cooccurrences et collocations fréquentes.

À l'usage, ces fiches CNC/XML s'avèrent plus particulièrement utiles pour le traitement (la rédaction des articles) de certaines catégories de mots, notamment des mots outils. En effet, dans ceux-ci, souvent hyper-fréquents, la redondance des cooccurents est particulièrement marquée et se prête bien non seulement en soi à une structuration mais, surtout, à la synthèse, le tout sans perte d'informations (ou, à tout le moins, avec une perte calculée comme minimale et perçue comme tolérable). Ici, toujours à l'usage, il est apparu préférable de travailler à partir d'une synthèse plutôt qu'avoir l'ensemble des contextes : il est en effet peu probable qu'un rédacteur ou un réviseur souhaite relire les 27 651 contextes du mot *avant*!

5.2. *Les limites des fiches CNC (BDTS-concordances)*

En ce qui concerne le traitement des mots autres que les mots outils, il semble que certaines nuances doivent être exprimées et que cette approche constitue un choix moins intéressant pour 2 principales raisons :

- La fiche CNC ne conserve pas de traces du travail fait sur les contextes ;
- Le rédacteur ne dispose pas de l'ensemble des contextes or l'expérience a montré que tous les rédacteurs souhaitaient voir l'*ensemble* des contextes, même si la fiche CNC existe.

Pour conserver son intérêt, il est apparu que le document synthèse de l'analyse des contextes de la BDTS que constitue la fiche CNC doit être élaboré par le rédacteur à qui est confié un mot aux fins de rédaction de l'article. Ce faisant, il dispose d'une vision globale du mot et est au courant des analyses réalisées sur les contextes, une condition qui est apparue difficilement négociable quand il s'agissait d'accorder sa confiance aux résultats de tris et à l'organisation des informations. Dans cette perspective, il est apparu que la spécialisation du travail n'était pas souhaitable puisque le rédacteur, en dépit de l'existence d'une fiche CNC réalisée par un autre que lui-même, ressentait le besoin de retourner dans la Banque pour faire son travail. Également, il est apparu que le retour différé sur un article donné (quelques mois plus tard, par exemple) pouvait même requérir de procéder à nouveau à l'analyse des contextes, la synthèse « inscrite » dans la fiche CNC ne satisfaisant pas nécessairement le réviseur.

En dernier lieu, l'examen des pratiques des rédacteurs a démontré que le travail dans l'éditeur XML dans lequel se déploient les DTD sous la forme de fiches informatisées n'était pas également aisé pour tous et que certains préféraient traiter manuellement les contextes dans un éditeur de texte. Mais ce qui est au cœur de la réorientation de la méthode de travail en ce qui concerne le traitement des contextes est beaucoup plus fondamental. C'est le résultat d'une « projection » de la tâche de rédaction des articles au moment du traitement des contextes des verbes, combinaison qui a marqué un changement progressif de perspective en faveur d'une approche d'analyse et de structuration des informations fondée sur la sémantique plutôt que sur la syntaxe et ce, pour tous les types de vocables, y compris les mots outils. Il s'agit là d'une façon de faire innovatrice dans le domaine de la lexicographie appliquée.

5.3. *Une perspective sémantique plutôt que syntaxique : la méthode « alternative » de traitement des contextes*

Lors du traitement des contextes des verbes notamment, plusieurs rédacteurs ont progressivement développé (et préféré employer) une autre méthode qui leur permettait, celle-là, non seulement de créer une synthèse de l'analyse des contextes, mais encore de laisser des traces du travail d'analyse ayant conduit à cette synthèse. Moins axée sur la réduction de la masse documentaire que sur la structuration de cette masse, la procédure s'est progressivement cristallisée autour d'un tri fondé sur la sémantique plutôt que sur la syntaxe. Cette méthode initialement « alternative » fait appel au dossier-sens et aux fichiers textuels de contextes (voir à ce sujet le point 2 et la Figure 1).

Il s'agit d'abord de procéder à l'extraction des paires sens-construction des dictionnaires (disponibles dans les fiches du Thésaurus Documentaire) et de créer un dossier-sens. L'étape suivante consiste à trier les contextes selon les paires contenues dans ce dossier et selon les spécificités de la Banque, façon de faire qui présente l'avantage de permettre l'enrichissement du dossier-sens par l'intégration des « nouveautés » identifiées dans la BDTS.

Il en résulte un nouveau fichier textuel dans lequel **tous** les contextes sont présents, mais regroupés par paires sens-construction (voir la Figure 3, plus bas). Ce fichier est complété par un second document, le dossier-sens lui-même, qui propose la synthèse de l'analyse : paires sens-construction avec fréquences associées, regroupées par blocs sémantiques, avec commentaires au besoin (voir le tableau ci-dessous). La complétude du tandem ainsi formé est telle qu'il redevient envisageable de spécialiser certains rédacteurs dans leur construction et ce, à la satisfaction de la majorité des autres rédacteurs qui peuvent alors se concentrer sur la tâche de rédaction des articles.

GRISER [fréquence = 52 après tri des contextes]

###[(rendre un peu ivre) GRISER QQN : A] [v. trans. dir.] **[5]**

###[(rendre un peu ivre) SE GRISER : S] [v. pron. réfléchi] **[3]**

###[(fig. enivrer de, repaître de) GRISER QQN DE : A] [v. trans. dir.] **[17]**

###[(fig. enivrer de, repaître de) ÊTRE GRISÉ DE, PAR : S] [voix passive] **[11]**

###[(fig. enivrer de, repaître de) GRISÉ : S] [part. passé adj.] **[2]**

###[(fig. enivrer de, repaître de) SE GRISER DE QQCH : A] [v. pron. réfléchi] **[13]**

###[(ternir) GRISER QQCH : S] [v. trans. dir.] **[1]**

###[IGNORÉS] **[2]**

Tableau 1. Exemple du contenu du dossier-sens pour le vocable « griser »

Il convient de remarquer qu'il est toujours également possible de consulter une fiche CNC qui aurait été construite pour avoir un aperçu des cooccurrents avant de commencer le tri des contextes; cependant, dans la plupart des cas, le fait d'avoir les contextes alignés sur le pipe « |* » permet de repérer ces cooccurrents rapidement, visuellement, ce qu'on ne pouvait faire avant d'avoir les contextes sous cette forme.

Comme il est possible de le constater dans la saisie d'écran (très fragmentaire) suivante, non seulement les différentes formes du vocable s'affichent et se regroupent sous le sens approprié, mais aussi le nombre de contextes, sous chacun, est-il rapporté. De plus, il est possible à l'aide de quelques commandes de retrier à volonté les informations.

```
                    ###[(rendre un peu ivre) GRISER QQN : A] [v. trans. dir.] [5]

inissait par me |*griser. Puis le lait s'ajoutant au lait, le pin#pon devenait de_plus en plus s
vin mousseux la |*grisaient plus_qu un demi#litre de whisky. Elle s'écria : - Ah, le Champagne,
e que le vin me |*grisait. Cela devenait grossier. En elle tout rayonnait de vivacité. Tout son
 en alcool, qui |*grise rapidement. ■CHAUD■ Vin riche en alcool ■COMPLET■ Vin bien équilibré, ré
'alcool l'avait |*grisé tout_de_même, le faisant pénétrer par moments dans une bulle de rêve où

                    ###[(rendre un peu ivre) SE GRISER : S] [v. pron. réfléchi] [3]

 en_train_de se |*griser de gin, comme un simple mortel. Le fabricant de conserves alimentaires
 timidement, se |*grisait tous les soirs. Les deux |P140 soeurs étaient veuves et jeunes. Les ma
morphinomane se |*grise de sa drogue. Un peu plus tard, il devint imaginatif. Il me confia, un s

                    ###[(fig. enivrer de, repaître de) GRISER QQN DE : A] [v. trans. dir.] [17]

enfant Qui vous |*grisa de mots, certain soir triomphant. Ce poème s'intitule « Victoire », mais
ouvait que nous |*griser. C'étaient des images d'un audacieux possible qui accaparaient la consc
e. On se laisse |*griser par le silence, rompu par le seul bruit des pagaies dans l'eau. Puis, c
n acheva de les |*griser en battant aux tempes et contre les joues de sa pulsion irrégulière. Gr
e de se laisser |*griser par ses succès remarquables de_la saison dernière. ■« Mon objectif pour
enaient, qui la |*grisaient du fond de_la glace. Plus tard, dans la soirée, alors_qu ils descenc
ents. Ce qui la |*grisait surtout, c'était dans une glace profonde_derrière le jeune homme, sa p
: cette idée me |*grisait. J'avais le sentiment de communier à la culture universelle. Je me ber
ent amputé, que |*grisent les musiques dépossédées et l'absence de terre où trouver le miel de_l
une auréole qui |*grise les foules et |P54 leur fait croire à l'invincibilité au moment même où
 / le danger le |*grise / m'man ça la avec l'escalier de sauvetage / il atteignait le toit / pas
le d'être Je me |*grise de voir er de toucher Je m'enflamme de chaque floraison Et chaque grain
nt. La lutte me |*grise et m'entraîne. Je ne suis pas né lutteur, toutefois c$est l'état de vie
 malgré tout le |*grise encore... la triste conviction qu'elle est troublante et jolie à faire t
ion de_l art le |*grise, il serait prêt à se dévouer, à accomplir des actes d'héroïsme pour sout
 l'espace, elle |*grise le cerveau humain d'un rêve tenace de libération, elle nous donne un oei
xquise qui vous |*grise, exactement comme l'air fameux de_la Veuve joyeuse qui est bien l'opéret

                    ###[(fig. enivrer de, repaître de) ÊTRE GRISÉ DE, PAR : S] [voix passive] [11]

sentir entouré, |*grisé de conquêtes, irrésistible séducteur. Sans conscience, il |P39 jouait à
me suis endormi |*Grisé du grand bonheur De c'qu'elle m'avait appris Hier, toute la journée Je r
le jeune homme, |*grisé par ces mots-là, fier de lui#même tandis que, par instants, la bourrasqu
```

Figure 3. La présentation des contextes triés par sens

5.4. *Intérêts et limites de la méthode « alternative »*

Cette méthode est apparue à la fois rapide, simple et efficace. Elle permet de garder une trace claire du travail d'analyse effectué et de conserver l'accès à tous les contextes. Ceux-ci sont maintenant également triés par paires construction-sens et regroupés en blocs sémantiques, ce qui constitue une avancée très intéressante et nous permet d'envisager, à moyen terme, la constitution d'un dictionnaire de cooccurrents sémantiquement groupés.

Avec, en main, la synthèse et l'ensemble des contextes triés, les rédacteurs éprouvent peu le besoin de retourner dans la Banque, sauf dans les cas où le contexte de 100 caractères sur lequel ils souhaitent se pencher particulièrement n'est pas suffisant ou pour puiser dans le corpus de citations. Elle peut également, par la suite, alléger le travail de révision des articles dans la mesure où il n'est pas nécessaire de refaire les analyses, mais seulement de les évaluer. Tous ces avantages nous indiquent qu'il y a lieu de favoriser cette méthode et de trouver les moyens de pallier sa seule carence (majeure, cependant) : les documents ainsi produits ne sont pas accessibles par la BDI, alors que cette intégration constitue l'une des raisons d'être de la plateforme informatique.

6. Conclusion

La BDTS-concordances constitue, en soi, un outil important et ce, à plusieurs titres. L'allègement et l'optimisation du traitement des contextes des mots outils et des autres mots

de fréquence très élevée identifiés dans la Banque de données textuelles est un impératif. Cependant, dans la pratique, il est apparu qu'avec leur forme actuelle, l'utilité des fiches développées selon le protocole initial n'est pas aussi grande que ce qui était anticipé lorsqu'il s'agit du traitement de vocables d'un autre type dans le cadre d'une tâche dictionnairique. En effet, les rédacteurs, s'ils désirent accéder aisément aux cooccurrents et aux constructions, réclament néanmoins la possibilité de voir la totalité des contextes. Il fallait donc se poser la question de l'intérêt actuel de la poursuite de la constitution de fiches CNC telles quelles, ce qui pouvait exiger une suspension de l'idéal de lisibilité et, surtout, d'accessibilité et de collaborativité tant locales (fiches consultables dans l'interface unique de la plateforme informatique) qu'étendues que permet l' « universalité » du langage de balisage structurel XML (ISO). Ces objectifs d'accessibilité et de collaborativité qui ont gouverné (et gouvernent encore) le développement de la plateforme intégrée propriétaire de FRANQUS ne devaient pas aller à l'encontre de notre conviction suivant laquelle il revient à l'informatique d'instrumenter une tâche et non à celle-ci de se plier à des contraintes informatiques, une perspective réductrice.

La lexicographie est une pratique qui doit absolument tenir compte des impératifs cognitifs de ceux-là mêmes qui en sont les artisans. Une solution possible au maintien de la **procédure** initialement prévue nous vient des rédacteurs eux-mêmes alors qu'ils énoncent qu'elle n'est réellement « intéressante » que lorsqu'ils prennent eux-mêmes en charge la construction de la fiche CNC du vocable dont la description leur est confiée. Dans cet esprit, il est apparu qu'il n'y avait pas lieu de spécialiser les rédacteurs par type de tâches (par exemple, rédaction des fiches CNC, rédaction des articles), mais plutôt de favoriser une approche dont l'homogénéité serait fondée et maximisée par une spécialisation thématique et/ou grammaticale plutôt (type de partie du discours).

L'ouverture avec laquelle la question a été abordée s'est traduite par des gains alors que les rédacteurs ont développé, parallèlement, une façon de faire permettant de combler leurs besoins tout en privilégiant une procédure fondée sur des critères de structuration sémantiques plutôt que syntaxiques. La voie est ouverte vers une BDTS-concordances « nouvelle et améliorée », plus riche en termes de contenu et porteuse de nouveaux développements, par exemple la constitution, à moyen terme, d'un dictionnaire des cooccurrents groupés de manière sémantique. Il s'agira maintenant pour nous de développer les modèles, les stratégies et les routines nécessaires à une prise en charge informatique correspondant à nos objectifs.

Références

Brunet Ét. *Hyperbase v. 5.4 : logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels*.

Fellbaum C. (1998). *WordNet, an Electronic Lexical Database.* MIT Press.

Gorcy G. (1990). Le Trésor de la langue française. Son originalité et les voies ouvertes pour son informatisation. In *CNRS/INaLF Dictionnairique et lexicographie. Autour d'un dictionnaire : le Trésor de la langue française, témoignages d'atelier et voies nouvelles*. Didier Érudition : 187-207.

Leroy-Turcan I. et Wooldridge R. (1998). *Quelques exemples des acquis de la base informatisée de la première édition du Dictionnaire de l'Académie française (1694).* Document hypertexte. http://www.chass.utoronto.ca/~wulfric/academie/acad1694/quebec298.htm.

Masson C.-É. (2001). *Le traitement des substantifs dans Le Robert sur CD-ROM : modélisation, formalisation et proposition méthodologique en vue de son informatisation*. Thèse de doctorat, Université de Sherbrooke.

Pruvost J. (2000). *Dictionnaires et nouvelles technologies.* Presses Universitaires de France.

Tutin A. et Wionet C. (1998). *Informatisation du Dictionnaire universel de Furetière revu par Basnage de Bauval (1702) : premier bilan.* Document hypertexte. http://www.chass.utoronto.ca/~wulfric/siehlda/dicta1998/tw_tab.htm.

University of Virginia. *TEI Guidelines for Electronic Text Encoding and Interchange*. Document hypertexte. http://etext.lib.virginia.edu/bin/tei-tocs?div=DIV2&id=DIEN.

Véronis J. et Ide N. (1996). Encodage des dictionnaires électroniques : problèmes et propositions de la TEI. In Piotrowski D. (Ed.), *Lexicographie et informatique. Autour de l'informatisation du Trésor de la langue française. Actes du Colloque international de Nancy*. Didier Érudition : 239-261.

Wooldridge R. *Baliser un texte, c'est le penser : le cas du Dictionnaire de l'Académie française.* Document hypertexte. http://www.chass.utoronto.ca/~wulfric/articles/gehlf597/index.html.

Les mots inconnus sont-ils des noms propres ?

Denis Maurel

LI (Université de Tours) – 64, avenue Jean-Portalis – 37200 Tours – France
denis.maurel@univ-tours.fr

Abstract

The voluminous textual data parsing is confronted with a specific class of words, the proper names, above all in case of business and strategic intelligence or information retrieval. The proper names are not often found in electronic dictionaries and one makes no distinction between them and unknown words. Meanwhile, instead of just declare to lump together proper names and unknown words, we present in this paper the result of a specific tagging experiment in every proper name and every unknown word of a newspaper issue.

Résumé

L'analyse de données textuelles volumineuses, surtout dans le cadre de la veille stratégique ou de la recherche d'information, est confrontée à une catégorie spécifique de mots, les *noms propres*, souvent absents des dictionnaires électroniques et, de ce fait, assimilés à des « Mots inconnus ». Cependant, au lieu de se contenter d'affirmations communes sur l'amalgame entre mots inconnus et noms propres, nous présentons ici les résultats d'une expérience de catégorisation de tous les mots inconnus et tous les noms propres d'un numéro du journal *Le Monde*.

Keywords: NLP, proper names, unknown words, tagging.

Mots-clés : TAL, noms propres, mots inconnus, étiquetage.

1. Introduction

L'analyse de données textuelles volumineuses, surtout dans le cadre de la veille stratégique ou de la recherche d'information, est confrontée à une catégorie spécifique de mots, les *noms propres*. Ceux-ci sont souvent absents des dictionnaires électroniques, alors qu'ils sont porteurs d'une riche sémantique tout en constituant, à eux seuls, plus de 10% des textes journalistiques (Coates-Stephens, 1993). En général donc, les noms propres rentrent dans la catégorie des « mots inconnus », c'est-à-dire non répertoriés par les différents dictionnaires de langues consultés, où se retrouvent les fautes d'orthographe, les abréviations, les dérivations peu communes, etc.

Par exemple, pour Ren et Perrault (1992), un mot inconnu commençant par une majuscule est considéré comme nom propre. Certains, moins expéditifs, utilisent des méthodes plus fines pour repérer les noms propres, souvent à l'aider de grammaires locales (Mac Donald, 1996 ; Mani et Richard MacMillan, 1996 ; Coates-Stephens, 1993) et proposent même parfois de classer les noms propres identifiés (Paik, *et al.*, 1996 ; Friburger et Maurel, 2001). En effet, la recherche et l'extraction d'informations ou l'aide à la traduction nécessitent de délimiter précisément les noms propres, de les catégoriser et même, parfois, de les relier entre eux, comme cela est proposé dans le projet *Prolex* (Maurel *et al.*, 1996) sur le traitement automatique des noms propres. Dans le cadre de ce projet, aujourd'hui multilingue, nous avons défini une typologie des noms propres et de leur relations (Grass *et al.*, 2002), que nous utiliserons ici pour présenter nos résultats.

Cependant, au lieu de se contenter d'affirmations communes sur l'amalgame entre mots inconnus et noms propres, nous avons tenté une expérience et nous avons catégorisé tous les mots inconnus et tous les noms propres d'un journal (*Le Monde*, daté du 15 janvier 1999), après l'utilisation d'un logiciel d'analyse lexicale, le logiciel *Intex* (Silberztein, 1993), dans sa version pour le français de septembre 2001.

2. Définir les noms propres

Avant de chercher à reconnaître des *noms propres*, il est nécessaire de définir cette catégorie linguistique. Mais la définition des noms propres est loin de faire l'unanimité parmi les linguistes. La définition classique est celle que donne, par exemple, le *Bon Usage* (Grevisse et Goosse, 1986 : 751) : « Le nom propre n'a pas de signification véritable, de définition ; il se rattache à ce qu'il désigne par un lien qui n'est pas sémantique, mais par une convention qui lui est particulière. » Elle est relayée par les propos de Gary-Prieur (1994 : 7) : « Alors que l'interprétation d'un nom commun ne met en jeu que la compétence lexicale, celle du nom propre requiert presque toujours une mise en relation avec le référent initial, qui mobilise des connaissances discursives. »

Cette définition correspond bien à des noms de personne (*Chirac*) ou de lieu (*Belgique*, *Bruxelles*). Mais elle nécessite d'être complétée afin de rendre compte de la réalité. C'est (Jonasson, 1994) qui distingue les noms propres « purs » (correspondant la définition de Grevisse) et les noms propres « descriptifs » qui résultent souvent de la composition d'un nom propre avec une expansion, comme *Jardin des Plantes* ou *Organisation mondiale de la Santé*. Certains semblent être des descriptions définies figées ou en cours de figement. La classe des noms propres purs est relativement fermée, alors que celle des noms propres descriptifs est ouverte à la création lexicale.

Nous prendrons donc comme définition celle de Jonasson (1994 : 21) qui déclare *nom propre* : « Toute expression associée dans la mémoire à long terme à un particulier en vertu d'un lien dénominatif conventionnel stable. » Cette définition, plus souple que la précédente, permet d'inclure les noms propres descriptifs, qu'il est convenu dans le monde du traitement automatique des langues, depuis les conférences *MUC*, d'appeler *entités nommées*[1] (Chinchor, 1997).

Souvent les linguistes mettent en avant le caractère de « désignateur rigide » du nom propre. Ainsi, selon Kripke (1980), un nom propre doit « désigner le même particulier dans tous les mondes possibles » et, selon Kleiber (1981 : 316), il n'est pas lié « aux situations passagères et aux propriétés accidentelles que peut connaître un particulier ». Cela devrait permettre de distinguer le nom propre (notamment descriptif) de la description définie, puisque le référent d'un nom propre peut évoluer sans que le nom propre utilisé ne soit changé.

Il est clair que *le président Chirac* n'est pas un nom propre, mais une description de la fonction politique exercée par l'homme qui a pour nom propre *Chirac*. Cependant, cette distinction n'est plus vraie pour les formes longues des noms de pays où la description est figée par des considérations politiques (Piton et Maurel 1997). Le dernier exemple en date est celui du *Zaïre* qui est devenu, suite à un coup d'état, la *République démocratique du Congo*. Peut-on dire que le référent (le pays africain désigné) n'est plus le même qu'avant ? Cela semble difficile à admettre. Et, dans une perspective de recherche d'information, il faudrait sans doute associer les deux termes.

[1] Pour être exact, les entités nommées comprennent non seulement les noms propres, mais aussi les dates et les mesures.

Enfin, il nous faut signaler l'importance de la métonymie des noms propres dans les textes journalistiques : *Paris*, *France*, *Français* peuvent désigner le même référent, le *gouvernement français*, et non leur propre référent. C'est ce qui nous a conduit, dans le projet *Prolex*, à envisager le nom propre comme une entité relationnelle (Maurel *et al*., 2000).

Enfin, en français, le critère formel le plus remarquable du nom propre, qui le distingue du nom commun ou de la description définie, est la majuscule. Celle-ci « indique le début d'une expression linguistique qui, d'un certain point de vue, constitue un tout » (Gary-Prieur, 1991 : 22). Mais cette majuscule n'apparaît pas sur tous les termes d'une entité nommée, ce qui pose le problème de la reconnaissance de sa limite droite (Friburger, 2002).

3. Le logiciel *Intex*

Le logiciel *Intex* est basé sur plusieurs bases de connaissance, hiérarchisées en quatre niveaux :

1. Les caractères, qui sont regroupés en trois classes : les chiffres, les lettres (définies dans un fichier d'alphabet) et les séparateurs (les autres caractères).
2. Les phrases, qui sont décrites par un transducteur à nombre fini d'états, chargé de découper le texte.
3. Les séquences de lettres (ou *mots simples*), qui sont répertoriés dans un dictionnaire de formes fléchies comportant 675 249 entrées, le *Delaf* (Courtois et Silberztein, 1990).
4. Les séquences de mots (ou *mots composés*), qui sont aussi répertoriés dans des dictionnaires, mais que nous n'utiliserons pas dans cette étude.

Après l'analyse lexicale par *Intex*, avec les bases 1, 2 et 3 ci-dessus, du journal *Le Monde* daté du 15 janvier 1999, nous avons obtenu les résultats suivants :

- 3 945 phrases.
- 114 217 éléments (dont 14 672 différents), répartis en séquences de lettres (86 409, dont 14 633 différentes), en chiffres (4 875) et en séquences de séparateurs (22 933, dont 29 différentes).
- 12 716 mots simples (présents dans le dictionnaire Delaf) et 1 917 mots inconnus.

Précisons ce que le logiciel *Intex* désigne par *mot inconnu* : il s'agit d'un mot formellement non présent dans le dictionnaire *Delaf*. Par exemple, la consultation du dictionnaire pour la phrase : « J'ai aperçu Jean, Pierre, Denis et Agnès. » donnera les résultats de la *Figure 1*.

Mots simples[2]	Mots inconnus
ai,avoir.V+z1:P1s aperçu,apercevoir.V+z1:Kms aperçu,aperçu.N+z1:ms et,et.CONJC+z1 jean,jean.N+z1:ms pierre,pierre.N+z1:fs pierre,pierrer.V+z2:P1s:P3s:S1s:S3s:Y2s	Agnès Denis

Figure 1. Les listes de mots après la consultation du dictionnaire

[2] Le *j'* a été reconnu lors de la préanalyse comme une élision non ambiguë (*j',je.PRO+PPV:1ms:1fs*), c'est pourquoi il n'apparaît pas ici.

Ainsi, à cette étape, *Pierre* ne sera pas un mot inconnu, à cause de *la pierre* et du verbe (rare[3]) *pierrer*, *Jean* non plus, à cause du pantalon, *le jean*, alors que *Agnès* et *Denis* le seront.

4. Les mots inconnus

Les mots inconnus de ce journal représentent 4 % de l'ensemble des mots simples du texte et 13 % des mots distincts (*Figure 2*).

Mots simples	Tous les mots		Les mots distincts	
Reconnus	82 524	96%	12 716	87%
Inconnus	3 885	4%	1 917	13%
Total	86 409	100%	14 633	100%

Figure 2. La répartition des mots simples du journal

Sur les 1 917 mots inconnus distincts, 243 ne sont pas des noms propres. Ceux-ci représentent donc 87 % des mots inconnus distincts. Cependant, sur les 243 mots inconnus distincts qui ne sont pas des noms propres, seul 23 commencent par une majuscule. En assimilant les 1 940 mots inconnus commençant par une majuscule à des noms propres, on ne se trompe que d'un peu plus de 1 %.

La principale source de mots inconnus (108, soit plus des deux cinquièmes) provient de la création lexicale :

- 21 abréviations : *Apocal*, *convalo*, *Corresp*, *éd*, *ème*, *gouv*, *MM*, *Mme*, *NDLR*, *rens*, *Tél*, etc.
- 11 chiffres romains[4] : *III*, *VIIIe*, *XIIe*, *XIXe*, *XVe*, *XVIe*, *XVIIe*, *XVIIIe*, *XXe*, *XXIe*, *XXs*.
- 15 dérivés de noms propres[5] : *ajaccien*, *belgitude*, *braudélienne*, *brejnévisme*, *cantabrique*, *delorien*, *khyal*, *lommoise*, *Newtonianisme*, *pivertisme*, *venézuélienne*, *Villeneuvoises*, *villiériste*, *wallisien*, *wolfienne*.
- 40 néologismes : *anthropogammamétrique*, *athéologique*, *autofiction*, *cacateux*, *europhorie*, *europrogressistes*, *germanité*, *médusante*, *métropolisation*, *milléniale*, *parentalité*, *pornocrate*, *rapporteure*, *remixage*, *surdemandés*, *surinterprétation*, etc.
- 5 onomatopées : *Epopopoï*, *Io*, *io*, *ito*, *ouba*.
- 14 sigles : *BB*, *CAFB*, *CMU*, *DG*, *HLM*, *NC*, *PDG*, *PIB*, *Ran*, *SBK*, *SDF*, *SMIC*, *TBP*, *ZEP*.
- 1 terme : *Erasmus*.

On totalise plus de la moitié des mots inconnus, si on ajoute à cette liste 31 emprunts :

- 4 mots latins : *exempla*, *Librium*, *tempora*, *vademecum*.
- 27 mots étrangers : *cyborgs*, *daddy*, *Euroland*, *joke*, *moderno*, *private*, *reals*, *sanpei*, *Seises*, *that*, *toros*, *völkisch*, etc.

Ensuite viennent les erreurs (53, un peu plus d'un cinquième) qui se répartissent entre :

[3] Le dictionnaire *Delaf* est divisé en trois couches, les mots courants (+z1), les mots rares (+z2) et les mots techniques (+z3).

[4] Le logiciel *Intex* permet l'utilisation d'un transducteur étiquetant les chiffres romains, mais nous n'avions pas sélectionner cette option, pour nous limiter aux seuls mots du dictionnaire *Delaf*.

[5] Dans le cadre du projet *Prolex*, nous avons étudié en détail la formation des gentilés (noms d'habitants), à partir d'un toponyme (Eggert *et al*., 1998 ; Eggert, 2002).

- 5 accents incorrects : *Nevadá*, *García*, *Martín*, *Vázquez*, *Louýs*.
- 9 fautes de frappe : *ccomptes*, *cespremiers*, *Cest*, *Cressson*, *Ilse*, *lerapproche*, *nérations*, etc.
- 38 fautes d'orthographe : *agrave*, *aseptie*, *automonie*, *cathéchiste*, *élevent*, *milllions*, *mutelles*, *secretaire*, *télévison*, *viendont*, etc.

Enfin, les 52 mots inconnus restants sont le résultat d'un mauvais paramétrage de la version d'*Intex* utilisée : les lettres majuscules accentuées ne figuraient pas dans le fichier d'alphabet et étaient donc assimilées à des séparateurs. La version actuelle du logiciel a corrigé ce défaut. La *Figure 3* présente cette répartition, secteur par secteur.

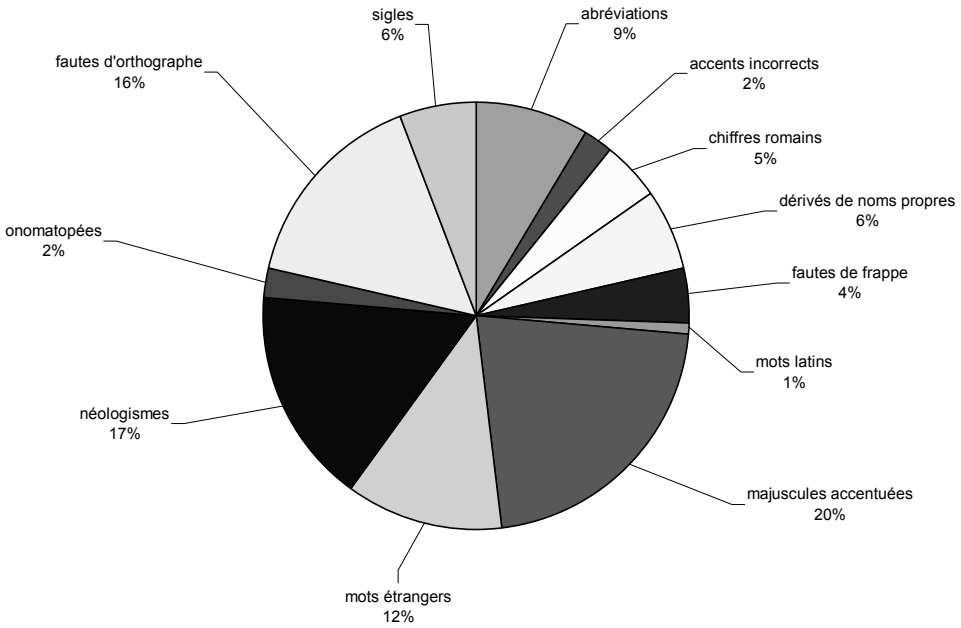


Figure 3. Les mots inconnus qui ne sont pas des noms propres

5. Les noms propres

Nous avons vu à la section 3 que les noms propres forment un ensemble de mots distinct de celui des mots inconnus, à cause de l'homographie de certains d'entre eux avec des noms communs. Mais, si nous suivons la définition de Jonasson (voir section 2), il y a plus : les noms propres sont souvent des mots composés, alors que les mots inconnus sont des mots simples. Le décompte des uns et des autres est donc radicalement différent, comme le montrent les exemples de la *Figure 4*.

Noms propres	Nombre de mots inconnus
Banque Centrale Européenne	0
Banco Real	1
Banca commerciale italiana	2[6]
Banca di Roma	3

Figure 4. Des noms propres et des mots inconnus

[6] Évidemment, les deux mots *Banco* et *commerciale* ne sont pas reconnus pour eux-même, mais ils possèdent des homographes en français.

Précisons le résultat donné à la section précédente : 243 mots inconnus distincts (sur 1 917) ne sont pas des noms propres, mais les 1 674 mots inconnus restant ne sont pas tous des noms propres purs, mais peuvent être simplement des parties de noms propres descriptifs. Pour connaître le nombre exact de noms propres du journal *Le Monde* du 15 janvier 1999, nous les avons effectivement dénombrés, par supervision humaine[7], pour arriver au chiffre de 1 419. Remarquons que, parmi ceux-ci, seuls 1 033 sont des mots simples. On constate donc qu'assimiler les 1 940 mots inconnus commençant par une majuscule avec des noms propres se révèle être très approximatif… Dans l'exemple de *Banco Real* (*Figure 4*), *Banco* n'est pas un mot inconnu et *Real* n'est pas un nom propre. C'est le mot composé *Banco Real* qui est intéressant à catégoriser.

Nous les avons ensuite classés en suivant notre typologie à deux niveaux (Grass *et al.*, 2002).

Après avoir étudié différentes classifications[8] concernant les noms propres, trop *ad hoc*, puis avoir envisagé d'utiliser le système des classes d'objets (Le Pesant et Mathieu-Colas, 1998), trop étendu pour être utilisé pour les seuls noms propres, nous avons préféré créer notre propre typologie, plus adaptée au traitement linguistique et notamment lexico-sémantique. Nous avons donc choisi de définir quatre *hypertypes*, qui reprennent en fait les traits classiques : les anthroponymes (trait *humain*), les toponymes (trait *locatif*), les ergonymes (trait *inanimé*) et les pragmonymes (trait *événement*) ; et, ensuite, vingt-neuf *types*, qui sont plus des *commentaires* portés sur les hypertypes, commentaires destinés principalement à l'aide à la traduction ou à la recherche d'information.

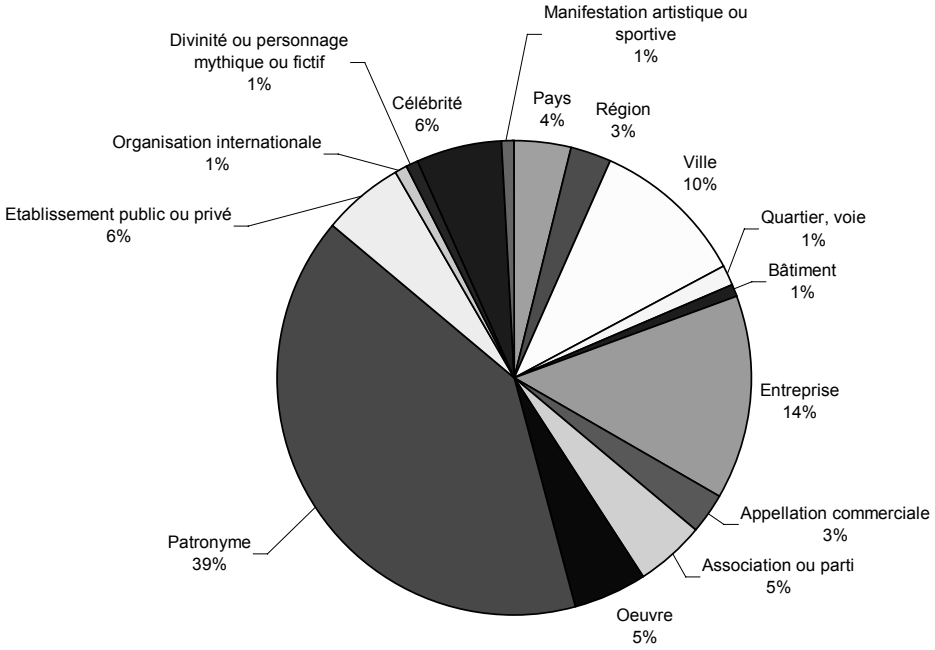


Figure 5. Les types des noms propres rencontrés

Parmi ces noms propres, comme on peut s'y attendre dans un quotidien, on trouve essentiellement des anthroponymes (1 010, presque les trois quarts), dont deux tiers de noms de personne et un tiers de noms collectifs. Viennent ensuite 284 toponymes (un cinquième de l'en-

[7] L'auteur remercie Caroline Ferré et Cécile Romanet pour leur aide dans le dépouillement des données.

[8] Issues de l'onomastique (Bauer, 1985), de MUC (Chinchor, 1997 ; Paik *et al.*, 1996), de l'économie (Zabeeh, 1968) et de la traduction (Ballard, 2001).

semble) et 108 ergonymes (un dixième), puis, très minoritaires, 17 pragmonymes. Le détail des types rencontrés est présenté sur la *Figure 5*, pour autant qu'ils représentent au moins 1 % du total ; de ce fait, quinze types seulement y sont présents (huit types sont sous-représentés et six sont absents du journal). Bien sûr, ces chiffres souffrent de la petitesse de notre corpus qui, dépouillé manuellement, ne pouvait être trop important.

Cette présence de nombreux noms propres composés (27%) justifie pleinement la mise en œuvre de techniques de reconnaissance adaptées à cette classe de mots. Dans le cadre du traitement automatique des langues, nous nous sommes attachés à développer des outils permettant de délimiter correctement les noms propres descriptifs.

Nous utilisons pour cela des cascades de transducteurs à nombre fini d'états (voir un exemple sur la *Figure 6*). Les résultats obtenus, cette fois-ci sur un corpus plus important (1 Mo de corpus journalistique) sont présentés dans (Friburger, 2002) et résumés sur la *Figure 7*.

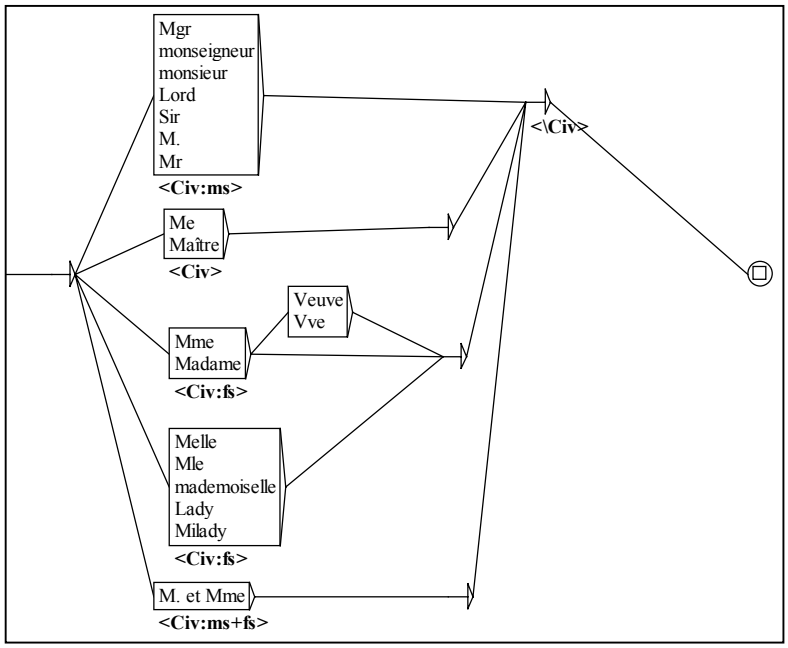


Figure 6. Un exemple de transducteur de la cascade : la reconnaissance des civilités

Reconnaissance automatique des noms propres Résultats obtenus sur 1 Mo de corpus journalistique	
Rappel	93,2
Précision	94,4

Figure 7. Les résultats obtenus par notre cascade

6. Conclusion

Notre étude, d'abord automatique, à l'aide du logiciel *Intex*, puis manuelle, d'un numéro du journal *Le Monde*, permet de vérifier précisément sur un exemple réel un certain nombre

d'affirmations souvent rencontrées, mais jamais chiffrées, concernant les mots inconnus, les noms propres et le rapport entre les deux.

Nous avons vu que, si les mots inconnus capitalisés sont effectivement presque toujours des parties de noms propres (à 1,2 % près), le décompte de ces derniers n'a aucun lien véritable avec les premiers, du fait que les noms propres descriptifs sont souvent composés en partie de noms communs. L'importance de la présence des noms propres a aussi été chiffrée précisément. Il n'est pas sûr qu'il soit pertinent de comparer ce nombre, qui est constitué, pour une grande part, de mots composés, à celui des mots simples du texte, mais, ce faisant, on obtiendrait alors 11 % de noms propres sur l'ensemble des mots distincts du journal. L'assimilation des mots inconnus commençant par une majuscule avec des noms propres est, elle aussi, très sujette à caution. La seule approximation acceptable est de les reconnaître comme partie de nom propre.

Références

Ballard M. (2001). *Le nom propre en traduction*. Ophrys.

Bauer G. (1985). *Namenkunde des Deutschen*. Germanistische Lehrbuchsammlung Band, vol. (21).

Chinchor N. (1997). *Muc-7 Named Entity Task Definition*. consultable sur le site http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html#appendices.

Coates-Stephens S. (1993). *The Analysis and Acquisition of Proper Names for the Understanding of Free Text*. Kluwer Academic Publishers.

Courtois B., Silberztein M. (1990). Dictionnaires électroniques du français. *Langues française*, vol. (87) : 11-22.

Eggert E. (2002). *La dérivation toponymes-gentilés en français : mise en évidence des régularités utilisables dans le cadre d'un traitement automatique*. Thèse de doctorat en linguistique.

Eggert E., Maurel D. et Belleil C. (1998). Allomorphies et supplétions dans la formation des gentilés. Application au traitement informatique. *Cahiers de Lexicologie*, vol. (73) : 167-179.

Friburger N. (2002), *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*. Thèse de doctorat en informatique, Université de Tours.

Friburger N. et Maurel D. (2001). Élaboration d'une cascade de transducteurs pour l'extraction de motifs : l'exemple des noms de personnes. In *Actes de la Huitième conférence annuelle sur le traitement automatique des langues naturelles (TALN 2001) :* 183-192.

Gary-Prieur M.-N. (1991). Le nom propre constitue-t-il une catégorie linguistique ? *Langue française*, vol. (92) : 4-25.

Gary-Prieur M.-N. (1994). *Grammaire du nom propre*. PUF.

Grass T., Maurel D. et Piton O. (2002). Description of a multilingual database of proper names. In *Proceedings of PorTal 2002*, LNCS.

Grevisse M. et Goosse A. (1986). *Le Bon Usage*. Duculot.

Jonasson K. (1994). *Le nom propre. Constructions et interprétations*. Duculot.

Kleiber G. (1981). *Problèmes de référence : descriptions définies et noms propres*. Klincksieck.

Kripke S. (1980). *Naming and Necessity.* Harvard University Press.

Le Pesant D. et Mathieu-Colas M. (1998). Introduction aux classes d'objets. *Langages*, vol. (131) : 6-33.

MacDonald D. (1996). Internal and external evidence in the identification and semantic categorisation of Proper Names. *Corpus Processing for Lexical Acquisition*. Massachussetts Institute of Technology : 21-39

Mani I. et Richard MacMillan T. (1996). Indentifying Unknown Proper Names in Newswire Text. *Corpus Processing for Lexical Acquisition*. Massachussetts Institute of Technology : 41-59.

Maurel D., Belleil C., Eggert E. et Piton O. (1996). Le projet PROLEX. In Séminaire *Représentations et Outils pour les Bases Lexicales, Morphologie Robuste de l'action Lexique du GDR-PRC CHM* : 164-175.

Maurel D., Piton O. et Eggert E. (2000). Les relations entre noms propres : lieux et habitants dans le projet *Prolex*. *Traitement automatique des langues*, vol. (41/1) : 623-641.

Paik W., Liddy E. D., Yu E. et McKenna M. (1996). Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval. *Corpus Processing for Lexical Acquisition*. Massachussetts Institute of Technology : 61-73

Piton O. et Maurel D. (1997). Le traitement informatique de la géographie politique internationale. In *Colloque Franche-Comté Traitement automatique des langues (FRACTAL 97)*, in *Bulag* (numéro spécial) : 321-328.

Ren X. et Perrault F. (1992). The typology of Unknown Words : An Experimental Study of Two Corpora. In *Proceedings of COLING 92*.

Silberztein M. (1993). *Dictionnaires électroniques et analyse automatique de textes - Le système INTEX*. Masson.

Zabeeh F. (1968). *What's in a Name, An Inquiry into the Semantics and Pragmatics of Proper Names*. Martinus Nijhoff.

Analyse logométrique de la cohabitation Chirac/Jospin (1997-2002). Explication de la défaite de Lionel Jospin à l'élection présidentielle de 2002

Damon Mayaffre

CNRS-ILF-Université de Nice
UMR 6039, Bases, Corpus et Langage (Nice)
mayaffre@unice.fr

Abstract

We propose to use the logometric method to study the corpus of the Chirac-Jospin cohabitation (1997-2002). Among other results, the method enables one to account for the Prime Minister's defeat : in 2002 a slow evolution of his discourse led to a break-away or divergence from his discourse of the previous years. As will be shown both by the figures measuring intertextual distances and by tree and factor analyses of the vocabulary or the grammatical tags, the socialist locutor has blurred the identity of his speech to the extent of no longer making himself heard.

Résumé

L'on se propose d'analyser, grâce à la méthode logométrique, le corpus de la cohabitation Jospin/Chirac (1997-2002). Cette étude, entre autres résultats, permet d'expliquer la défaite du Premier ministre : à la suite d'un processus linguistico-politique, le discours qu'il tient en 2002 apparaît en rupture avec celui qu'il tenait les années précédentes. Comme le montreront le calcul des distances intertextuelles, les analyses factorielles et arborées réalisées sur le vocabulaire et les codes grammaticaux, le locuteur socialiste brouille l'identité de son discours jusqu'à ne plus se faire entendre.

Mots-clés : logométrie, distance intertextuelle, analyse arborée, individuation socio-linguistique, discours politique, Chirac/Jospin.

1. Introduction

Les périodes de cohabitation sont toujours des moments particuliers de la vie politique française. Mais la cohabitation Chirac/Jospin est exceptionnelle à double titre : d'abord par sa longueur (5 ans), ensuite par la situation originale du président, qui sans avoir souhaité explicitement la division de l'exécutif français l'a directement provoquée par la dissolution de 1997. Contrairement aux cohabitations Mitterrand/Chirac (1986-1988) et Mitterrand/Balladur (1993-1995), la cohabitation Chirac/Jospin (1997-2002) n'est pas une courte parenthèse politique imposée par le calendrier, mais un moment prolongé, « choisi » par les deux parties, durant lequel la France, non sans schizophrénie politique, apprend à parler à deux voix.

C'est cette polyphonie politique que nous voulons mettre à l'épreuve en étudiant les discours de Chirac et Jospin grâce aux traitements logométriques et au logiciel *Hyperbase*.

Corpus

Le corpus étudié est comparatif (conditions du traitement logométrique) : il est contrastif (Chirac *vs.* Jospin) et diachronique (1997-2002). Il est composé de tous les discours que le président et le Premier ministre ont prononcés ou écrits en France entre 1997 et 2002 soit quelque 2100 discours, 2.400.000 mots, 140.545 vocables (tableau 1).

Textes	Nombre de discours	Nombres d'occurrences	Nombre de vocables différents
Chirac1997	72	104.553	8902
Chirac1998	102	183.820	11.910
Chirac1999	110	176.352	10.989
Chirac2000	129	251.570	12.917
Chirac2001	122	219.236	12.110
Chirac2002	69	209.694	11.638
Total partiel (1)	*604*	*1.145.225*	
Jospin1997	68	56.229	5.935
Jospin1998	98	190.455	11.736
Jospin1999	107	225.158	12.915
Jospin2000	106	300.308	15.035
Jospin2001	113	276.747	14.077
Jospin2002	77	227.700	12.381
Total partiel (2)	*569*	*1.276.597*	
Total	1173	2.421.822	34.121

Tableau 1. Le corpus de la cohabitation Chirac/Jospin (1997-2002)

Méthode

Nous entendons par traitement logométrique d'une part la batterie d'outils documentaires et statistiques aujourd'hui disponibles sur des logiciels comme *Hyperbase* ou *Lexico* (spécificités, AFC, analyse arborée, calcul des distances, etc.), et d'autres part le fait que ces outils s'appliquent à tous les niveaux linguistiques (mots, lemmes, codes grammaticaux, enchaînements syntaxiques).

Par là cette contribution espère participer à l'amélioration de l'ADT à double titre, d'une part en comparant la performance des différents outils : l'analyse arborée notamment sera avantageusement mise en parallèle avec l'analyse factorielle. D'autre part, en comparant les niveaux de l'analyse linguistique : du point de vue de la distance entre les textes, les analyses sur les formes graphiques, les lemmes ou les fonctions grammaticales aboutissent à des résultats voisins, comme si un texte formait un tout linguistique aux dimensions multiples mais à l'unité remarquable.

Problématique socio-linguistique

Face à un corps de discours segmenté politiquement et chronologiquement, deux types de questionnements peuvent être formulés. Premièrement y a-t-il une identité linguistique chiraquienne et une identité linguistique jospinienne ? Les deux hommes tiennent-ils le même langage ou bien des signes linguistiques forts et originaux marquent-ils leurs discours respectifs ? Cette interrogation prolonge notre réflexion sur l'individuation socio-linguistique des locuteurs politiques menée lors des dernières JADT à St. Malo (Mayaffre, 2002).

Deuxièmement, quel est le rythme et l'ampleur de l'évolution chronologique dans le corpus ? Les discours sont-ils identiques à eux-mêmes entre 1997 et 2002 ou changent-ils au fil des ans de manière continue et/ou chaotique ? Comment, finalement, évolue le parcours des deux hommes sous la cohabitation ?

Surtout, les deux types de questionnements peuvent être croisés. Quel est le facteur qui apparaît le plus prégnant dans le corpus ? Le facteur politique ou le facteur chronologique ? L'identité politico-linguistique de Chirac et Jospin résiste-t-elle aux contraintes imposées par le temps ? Les discours de Chirac — quelle que soit la conjoncture — s'opposent-ils toujours à ceux de Jospin ? Ou, au contraire, le temps, avec les thèmes qu'il charrie, commande-t-il les discours pour rapprocher les 2 hommes ?

Par là, le dénouement de 2002 et l'humiliante défaite de Jospin pourront être illustrés : à la suite d'un processus linguistico-politique, le discours que le Premier ministre tient en 2002 apparaît en rupture avec celui qu'il tenait les années précédentes, le locuteur socialiste brouille l'identité de son discours jusqu'à le rendre inaudible.

2. Identités discursives : deux hommes, deux statuts, deux discours

L'objet principal de cette contribution n'est pas là. Mais posons que Chirac et Jospin ont une identité linguistique différente[1]. Plusieurs raisons — qu'il serait vain ici de vouloir démêler — expliquent cette individualisation linguiste : les deux principales étant le statut des locuteurs (président/Premier ministre) et la culture politique des deux hommes (droite/gauche).

Les spécificités lexicales des deux hommes sont en tout cas éloquentes (tableau 2).

Spécificités de Chirac (1997-2002)	Spécificités de Jospin (1997-2002)
Notre (+29)	Gouvernement (+32)
Tout (+26)	Ce (déterminant) (+26)
Vous (+24)	Député (+24)
Naturellement (+21)	Corse (nom) (+23)
Monde (+20)	Franc (nom) (+20)
France (+20)	Ministre (+19)
Votre (20)	Politique (nom) (+19)
Ce (pronom) (+19)	Loi (+19)
Et (+19)	1997 (+18)
Son (+17	Emploi (+17)

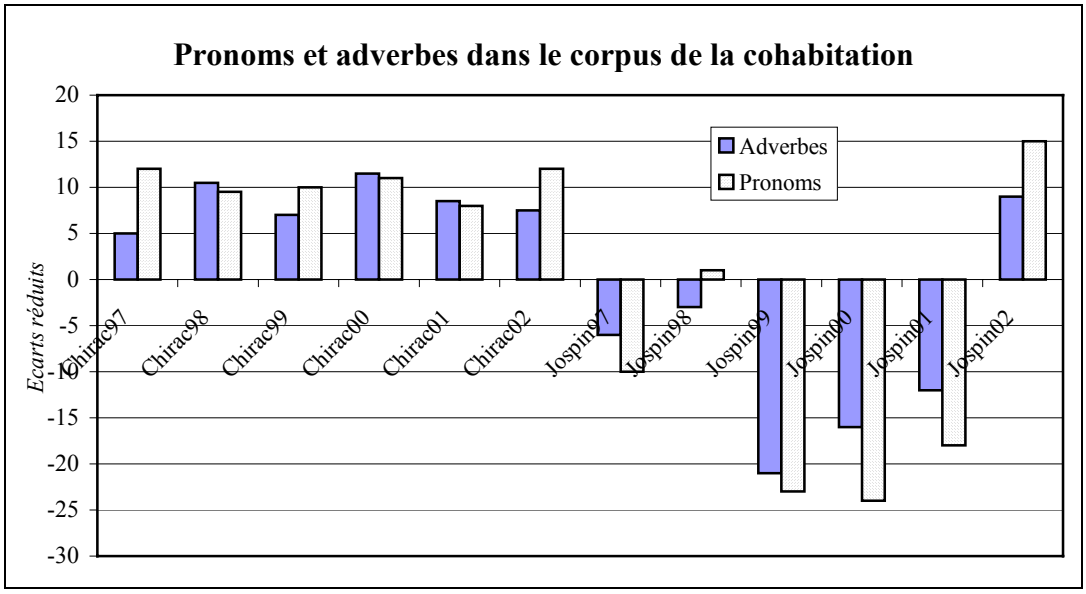
Tableau 2. Spécificités comparées de Chirac et Jospin (1997-2002)

Les spécificités de Chirac (ici les 10 plus importantes) relèvent — à l'exception de « France » et « monde » — de la mise en forme du discours. Il s'agit de mots-outils sans charge politique : « notre », « tout », « vous », « naturellement » etc. Les spécificités de Jospin sont, elles, de l'ordre du contenu politique du discours, qu'il s'agisse de termes institutionnels de la République ou de mots fortement estampillés de la langue politique : « gouvernement », « député », « loi », « emploi », « politique ». Du reste, la vacuité politico-linguistique est une

[1] Pour le développement qui suit cf. Mayaffre (2004 : chapitre 9).

des caractéristiques du discours de Chirac, tant au regard de Jospin que du discours d'un de Gaulle ou d'un Pompidou (Mayaffre, 2004 : Section I et II).

Grammaticalement, cela a une traduction nette et tout aussi explicite. Chirac (par rapport à Jospin) sur-utilise avant tout les adverbes ou les pronoms (graphique 1) et sous-utilise les noms porteurs de sens ou les adjectifs.



Graphique 1. Les pronoms et Les adverbes dans le corpus Cohabitation (1997-2002)[2]

En réalité, le discours présidentiel chiraquien ne dit pas grand chose, mais est fortement modalisé. Le discours de Jospin essaye de véhiculer un message quitte à paraître technique ou impersonnel (Mayaffre, 2004).

	C.97	C.98	C.99	C.00	C.01	C.02	J.97	J.98	J.99	J.00	J.01	J.02
C.97	0	0,110	0,116	0,113	0,124	0,145	0,197	0,167	0,183	0,179	0,173	0,168
C.98	0,110	0	0,089	0,094	0,098	0,119	0,182	0,148	0,168	0,167	0,162	0,151
C.99	0,116	0,089	0	0,088	0,092	0,123	0,180	0,151	0,160	0,163	0,160	0,155
C.00	0,113	0,094	0,088	0	0,083	0,118	0,171	0,142	0,155	0,150	0,147	0,146
C.01	0,124	0,098	0,092	0,083	0	0,107	0,166	0,143	0,150	0,150	0,139	0,148
C.02	0,145	0,119	0,123	0,118	0,107	0	0,178	0,150	0,168	0,168	0,161	0,139
J.97	0,197	0,182	0,180	0,171	0,166	0,178	0	0,133	0,123	0,124	0,126	0,175
J.98	0,167	0,148	0,151	0,142	0,143	0,150	0,133	0	0,114	0,118	0,113	0,131
J.99	0,183	0,168	0,160	0,155	0,150	0,168	0,123	0,114	0	0,094	0,096	0,158
J.00	0,179	0,167	0,163	0,150	0,150	0,168	0,124	0,118	0,094	0	0,092	0,152
J.01	0,173	0,162	0,160	0,147	0,139	0,161	0,126	0,113	0,096	0,092	0	0,144
J.02	0,168	0,151	0,155	0,146	0,148	0,139	0,175	0,131	0,158	0,152	0,144	0

Tableau 3.
Indices des distances entre textes selon le vocabulaire dans le corpus Cohabitation (1997-2002)

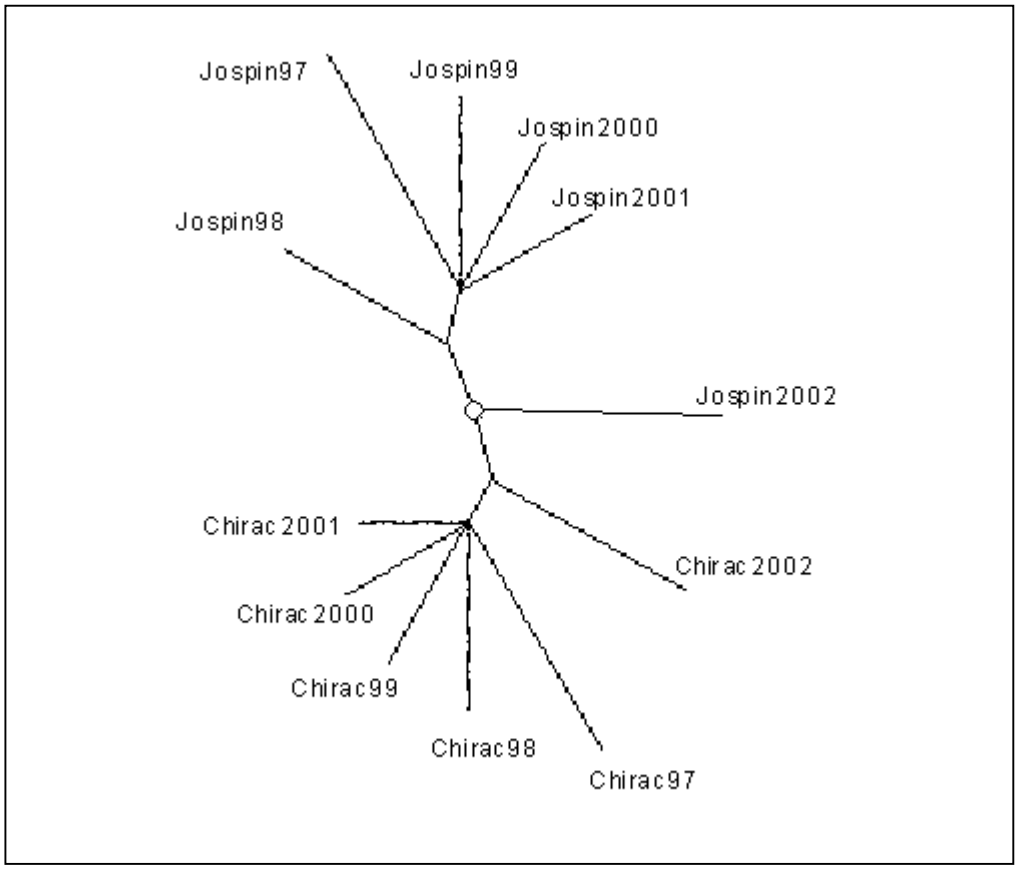
[2] La particularité de Jospin2002, évidente sur le graphique, fera l'objet de notre attention plus bas.

3. Le rythme de la cohabitation

Les identités respectives de Chirac et Jospin ainsi établies à grands traits, il convient de les mettre à l'épreuve du temps à l'image du dernier graphique. Le traitement diachronique sera ici global grâce à l'outil de synthèse que constitue l'analyse arborée.

Un indice de distance intertextuelle, calculé selon la méthode Labbé (2003), sur l'ensemble du vocabulaire employé dans les discours, a été établi pour aboutir au tableau matriciel suivant (tableau 3).

Une analyse arborée, selon la méthode Luong (1988 ; Barthélemy et Luong, 1998), qui allie techniques classificatoires et topologie mathématique[3], donne une représentation fine et précise de cette grande matrice (graphique 2).



Graphique 2. analyse arborée selon le vocabulaire du corpus Cohabitation (1997-2002)

À partir d'un tel arbre, trois conclusions socio-linguistiques peuvent être développées.

D'abord, il existe une individuation linguistique réelle entre Chirac et Jospin. L'algorithme dessine nettement deux branches opposées. Sur l'une se distinguent toutes les ramifications de Chirac, sur l'autre toutes les ramifications de Jospin. (On enregistre une seule exception, nous y reviendrons, le discours de Jospin 2002 qui forme à lui tout seul une branche indépendante.) Le temps ne perturbe que très peu cette dichotomie politique. Après beaucoup d'autres expériences, avec des outils variés, sur des corpus politiques divers, nous pouvons à nouveau affirmer que le facteur individuel et politique semble toujours prépondérant dans le discours

[3] L'analyse arborée a déjà fait l'objet de présentation aux JADT : succincte et pragmatique dans Luong et Mellet (1995) ; détaillée et technique dans Barthélemy et Luong (1998).

politique. Lorsque l'on croise plusieurs facteurs — l'identité politique du locuteur (sa personnalité et son statut), la chronologie, etc. — c'est toujours l'identité du locuteur — son appartenance politique et son statut — qui apparaît primordiale (Mayaffre, 2002).

Ensuite, le clivage entre Chirac et Jospin est réel mais *tend à se réduire* au fil de la cohabitation. Sur chacune des branches, c'est-à-dire localement sur l'arbre, le graphe se montre très précis notamment dans la description de la chronologie et de l'évolution du discours durant la période.

Il est très intéressant de remarquer comment le discours de Chirac 2002, par une sous-branche particulière, change en partie de nature et cherche à gommer son identité chiraquienne originelle. Il ne fait aucun doute que la campagne électorale se fait sur des thèmes jusqu'alors inexploités, avec d'autres mots que ceux des discours antérieurs. Pour des raisons électorales, Chirac semble inventer, en 2002, un autre discours. Néanmoins — et la précision prendra tout son sens plus loin —, ce discours de campagne en grande partie original reste clairement affilié au *parler* de Jacques Chirac : il reste, sur l'arbre, sur la même branche maîtresse chiraquienne avec les autres discours du président.

D'ailleurs, chez Chirac, cette évolution très marquée et un peu exceptionnelle du discours de 2002 ne doit pas nous masquer l'évolution chronologique progressive et cohérente de l'ensemble du discours. Sur l'arbre, dans le bouquet des textes de Chirac entre 1997 et 2001, on remarque que la distance (c'est-à-dire l'originalité) tend à se réduire au fil du temps : les ramifications sont en effet de plus en plus courtes, au fur et à mesure que s'égrainent les années. En d'autres termes — et l'observation attentive des indices du tableau 2 peut le confirmer — le discours de Chirac en 1997 est le plus éloigné du discours moyen (c'est-à-dire en grande partie, par effet de contraste dans notre corpus, du discours de Jospin), puis en 1998 le président perd en partie de son originalité comme l'atteste une ramification plus courte, en 1999, il en perd encore davantage, etc.

Du coté Jospin la même tendance est observée. À mesure que le temps s'écoule, Jospin se rapproche d'un discours moyen (c'est-à-dire dans notre corpus du discours chiraquien). Une seule exception perturbe légèrement ce bel ordonnancement chronologique : le discours de 1998 est très précocement édulcoré se démarquant de ses frères pour se rapprocher du centre topologique de l'arbre[4]. Mais au-delà de cette exception, c'est le discours de 1997 qui est le plus typé ou le plus marginal (ramification la plus longue), puis celui de 1999, puis celui de 2000, etc.

C'est pourquoi la conclusion est évidente : les deux hommes déteignent l'un sur l'autre, et leurs discours tendent à se rapprocher. L'effet émollient de la cohabitation où le clivage gauche/droite est estompé est ainsi illustré. Sans jamais tenir le même discours, Jospin et Chirac perdent petit à petit, entre 1997 et 2002, leur identité propre pour employer de plus en plus un vocabulaire commun.

4. La défaite discursive de Jospin en 2002

Mais l'intérêt de l'arbre présenté se trouve ailleurs, dans la situation particulière du texte Jospin2002. Le discours de Jospin n'évolue pas seulement mais transmue durant l'année 2002 pour devenir méconnaissable. Que les discours des deux hommes tendent à se rapprocher a déjà été dit. Que, symétriquement à Jospin, le discours de campagne de Chirac soit particulier

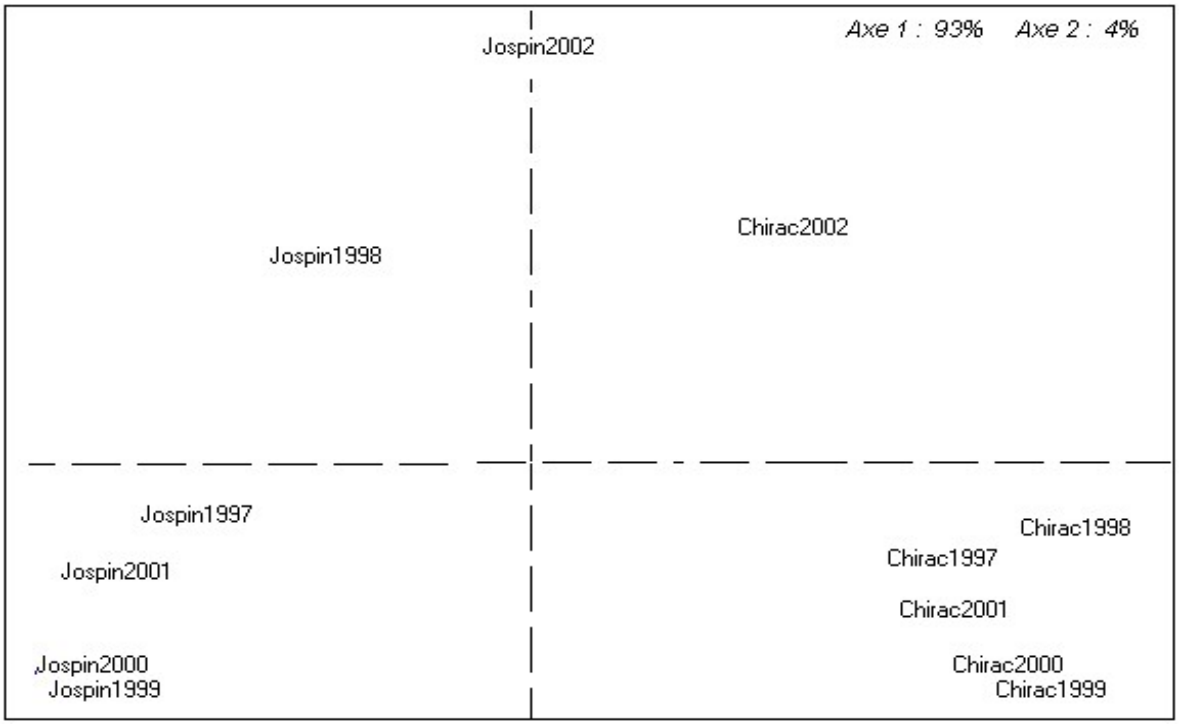
[4] Une explication de la particularité de l'année 1998 durant la gouvernance de Lionel Jospin est donnée dans Mayaffre (2004).

et opportuniste a déjà été souligné. Mais la localisation de Jospin2002 dépasse tous ces constats pour expliquer la défaite du Premier ministre aux élections présidentielles.

Il ne s'agit plus d'une évolution du discours mais d'une rupture linguistique étonnante. Dans le clivage linguistico-politique Jospin / Chirac dont elle donne une belle représentation, l'analyse arborée n'arrive plus à reconnaître et *à classer* Jospin2002. Le discours de campagne du Premier ministre se trouve aussi éloigné du discours de Chirac que des siens propres durant les années précédentes. En 2002 Jospin tient un discours à part, original, mutant, qui ne *s'apparente pas* à celui qu'il a tenu précédemment[5]. La personnalité linguistique (sinon politique) de l'émetteur, que les Français avaient appris à reconnaître durant 5 ans, est totalement brouillée. En termes rapides, Jospin se met subitement, comme Chirac, à utiliser des adverbes et des pronoms et à négliger les noms surtout lorsqu'ils sont politiques (Mayaffre, 2004). Le graphique 1, produit plus haut, n'est qu'une illustration générale du phénomène.

Pour cette raison, le discours du candidat Jospin devient inaudible particulièrement pour l'électorat de gauche, habitué à un autre discours. Est-ce la faute des conseillers en communication ou des plumes du candidat ? Est-ce la faute des médias et de ses adversaires qui ont su entraîner Jospin sur des thématiques qui lui étaient étrangères ? En tout cas les mots que le Premier ministre candidat se met alors à utiliser ne lui appartiennent pas. En 2002, Jospin vocalise un discours étranger, « inidentifiable ». Dans ces conditions, il avait peu de chances de se faire entendre.

Pour conclure, l'analyse est confirmée par une AFC issue du traitement des données lexicales (graphique 5). Ici l'axe 1 (horizontal) et l'axe 2 (verticale) ont été croisés.



Graphique 3. AFC du corpus Cohabitation (1997-2002) selon le vocabulaire

[5] On remarquera de manière symptomatique que le texte le plus éloigné (0.175) du texte Jospin2002 n'est pas un texte de Chirac mais le texte Jospin1997 : dans le corpus, les discours qui ressemblent le moins aux discours de Jospin 2002 sont les discours de Jospin tenus 5 ans auparavant.

Le schème général de l'AFC est identique à celui de l'arbre. Les conclusions linguistico-politiques, en trois points, sont confirmées : il existe bien une identité lexicale chiraquienne *versus* jospinienne, apparente sur l'axe 1 (axe horizontal recueillant 93% de l'information) ; les discours de Chirac2002 ou Jospin1998 apparaissent comme particuliers dans leur famille respective (particulièrement sur l'axe 2) ; enfin Jospin2002 n'est plus seulement particulier mais totalement a-typique, exactement central sur l'axe 1 et excentré sur l'axe 2.

Nous noterons enfin, à titre méthodologique, que localement sur l'AFC, les concentrés de points (Chirac1997, Chirac1998, Chirac1999, Chirac2000, Chirac2001 et Jospin1997, Jospin1999, Jospin2000, Jospin2001) sont peu précis et difficiles à interpréter sans la consultation de l'axe 3 voire de l'axe 4. L'analyse arborée, elle, nous avait pourtant permis de constater, du premier coup d'œil, l'évolution qui avait cours durant la cohabitation, en soulignant — par la dimension de plus en plus courte au fil des années des ramures menant aux textes — comment les discours du président et du Premier ministre perdaient, dans la durée, de leur originalité respective pour tendre vers leur confusion.

Références

Barthélemy J.-P. et Luong X. (1998). Représenter les données textuelles par les arbres. In *Actes des JADT 1998* : 49-73.

Brunet Ét. (1988). Une mesure de la distance intertextuelle : la connexion lexicale. *Revue informatique et statistique dans les sciences humaines*, *Le nombre et le texte, hommage à Etienne Evrard* : 81-116.

Brunet Ét. (2002). Le lemme comme on l'aime. In *Actes des JADT 2002* : 221-233.

Brunet Ét. (2003). Peut-on mesurer la distance entre deux textes ? *Corpus*, vol. (2) : 47-70.

Corpus (2003). Vol. (2), *La distance intertextuelle*.

Labbé C. et Labbé D. (2003). La distance intertextuelle. *Corpus*, vol. (2) : 95-118.

Labbé D. et Hubbert P. (1998). La connexion des vocabulaires. In Actes des *JADT 1998* : 361-370.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Xuan L. (1988). *Méthodes d'analyse arborée. Algorithmes, applications*, Thèse d'Etat, Paris V.

Luong X. (1994). « L'analyse arborée des données textuelles : mode d'emploi. *Travaux du cercle linguistique de Nice*, vol. (16) : 25-43.

Luong X. et Mellet S. (1995). Le calculs multidimensionnels au service de l'analyse syntaxique diachronique. In *Actes des JADT 1995* : 281-289.

Luong X. et Mellet S. (2003). Mesures de distance grammaticale entre les textes. *Corpus*, vol. (2) : 141-166.

Mayaffre D. (2002). Discours politique, genres et individualisation socio-linguistique. In *Actes des JADT 2002* : 517-529.

Mayaffre D. et Luong X. (2003). Arbres et généalogie politique. Représentation arborée du discours de Jacques Chirac (1995-2002). *Histoire et Mesure*, vol. (XVIII-3/4).

Mayaffre D. (2004), *Jacques Chirac (1995-2003) : le Président de la parole*. Champion, Lettres numériques. (A paraître).

Pincemin B. (2002). Similarités texte-textes. Expérience d'une application de diffusion ciblée et propositions. In *Matemáticas y Tratamiento de Corpus* (s. n.). Logroño : Fundacion San Millán de la Cogolla : 35-52.

Le texte dans tous ses états
Philosophie d'encodage du projet Khartês

Nicolas Mazziotta

Université de Liège – 4000 – Belgique
nicolas.mazziotta@ulg.ac.be

Abstract

When editing Medieval sources, one always encounters the same dilemma : how to make the text readable nowadays without spoiling its testimony ? « Cleaning » the text from all that makes it difficult to read for someone erases marks that would look valuable from another point of view. Hence, there is an unavoidable gap between the source and the edition. The *Khartês* project (which intends to edit and study original Old French charters written in Liège before 1300) has encountered this problem. Traditional « solutions » were inadequate and computers opened an interresting path. However, *Khartês* had to think about creating an encoding scheme that would suit its aim and remain easy to use. This is the way KML (the *Khartês Markup Language*, an XML-conformant language) was born, together with a small set of handy encoding rules. A wide range of applications are made possible. For instance, thanks to the way the *corpus* has been created, it has been possible to search a particular document for *marked letters* (which look like our uppercase letters) and to study them.

Résumé

Éditer un document médiéval pour que celui-ci soit appréhendable à l'époque actuelle n'est pas une mince affaire. Constamment tiraillé entre le respect du document et la volonté de le rendre accessible, l'éditeur en est réduit à contenter un type de public en particulier au détriment d'un autre. Son travail se solde toujours par une importante distance entre le support original et l'édition moderne ; c'est inévitable. Ce problème, *Khartês* (projet liégeois d'édition et d'étude linguistique de chartes originales françaises antérieures à 1300) l'a rencontré. Les « solutions » qu'offraient la perspective traditionnelle étaient insuffisantes, et l'ordinateur ouvrait des voies intéressantes. Néanmoins, il a fallu réfléchir à une manière de faire qui convienne tout à fait à notre propos tout en suivant la philosophie d'économie que les contingences matérielles nous obligeaient à suivre. Ce qui ressort de cette expérience, c'est un système de balisage relativement simple baptisé *Khartês Markup Language* (KML, qui est conforme à XML) et une petite série de règles d'encodage simples et légères. Les applications sont multiples. Par exemple, grâce à cette formalisation textuelle, nous avons pu mettre à l'épreuve le corpus et l'interroger au sujet de la répartition des *lettres marquées* (qui ressemblent à nos capitales actuelles) dans un document particulier.

Mots-clés : *représentations textuelles, chartes médiévales, XML, fatware.*

1. Introduction

Signalons pour commencer que notre contribution fait partie des recherches menées dans le cadre du projet *Khartês* — projet d'édition et d'étude linguistique des chartes originales rédigées en français en Wallonie avant 1300. Nous avons ici pour but d'exposer le principe fondamental qui guide l'encodage que nous avons employé pour la constitution du corpus du projet et son utilité dans le cadre d'études spécifiques. Après avoir brièvement présenté ce dernier (2) ainsi que ses objectifs (3) et situé notre travail personnel à l'intérieur de cette entreprise, nous exposerons les difficultés qu'a posées le passage du texte la formalisation des documents (4) et les solutions que nous avons retenues (5). Nous terminerons par un exemple d'application (6) avant de conclure (7).

2. Projet *Khartês* : présentation

Khartês est un projet d'édition et d'étude des chartes originales rédigées en langue française en Wallonie avant 1300. Il a débuté en 1998 à l'initiative de Marie-Guy Boutier (professeur à l'Université de Liège)[1] et trouve sa place dans la vaste entreprise d'édition que constitue la collection des *Documents linguistiques de la France*.

La collection des *Documents linguistiques de la France* est inaugurée en 1974, avec l'édition des documents originaux conservés dans le département de Haute-Marne, préparée par Jean-Gabriel Gigot sous la direction de J. Monfrin (Gigot, 1974). Il s'agit de la première pièce d'un vaste projet destiné à réunir les textes non littéraires rédigés en langue vulgaire dans le domaine français. La collection devait initialement comprendre trois séries couvrant les domaines d'oïl (série française) et d'oc, mais aussi le domaine du franco-provençal. Il y était notamment prévu d'éditer des chartes originales conservées dans les différents fonds d'archives de la France (la série franco-provençale est un recueil d'éditions de textes dialectaux non littéraires).

En 1984, parait le premier volume des *Documents linguistiques de la Belgique romane*, série parallèle à la série française des *Documents linguistiques de la France* et qui comprendra finalement trois volumes distincts : un volume rassemblant les chartes du Hainaut (Ruelle, 1984), un autre pour les chartes flamandes (Mantou, 1987) et un troisième, encore manquant, dédié au chartes de Wallonie. Répondant à l'invitation que lui fit personnellement Jacques Monfrin en 1994, M.-G. Boutier entreprend le projet *Khartês*, destiné dans un premier temps à fournir les matériaux nécessaires à l'élaboration du tome trois des *Documents linguistiques de la Belgique romane*, et donc à couvrir la Wallonie (provinces de Liège, de Namur et de Luxembourg). Le corpus étudié dans le cadre du projet *Khartês* sera constitué pour commencer des chartes originales conservées dans les provinces de Liège et de Namur. Le travail a débuté par l'édition de celles qui sont conservées aux Archives de l'État à Liège. Toutes les chartes originales antérieures à ca 1291 seront étudiées.

3. Objectifs

Pour commencer, voyons concrètement à quoi ressemble une charte (v. fig. 1, p. suivante). *Khartês* se pose d'emblée comme objectif d'éditer ces documents pour qu'ils servent de matériaux pour des recherches linguistiques ; le projet vise donc à « donner un texte à lire »[2] à des linguistes.

Or il se fait que les documents tels que celui que nous venons de montrer ont longtemps occupé une communauté de chercheurs qui n'étaient pas linguistes : les diplomatistes. En matière de typologie textuelle, les chartes appartiennent en effet toutes à une même grande famille[3], celle des textes diplomatiques. Indépendamment du public ciblé, les textes que *Khartês* édite doivent recevoir l'habillage traditionnel qui convient à ce type de texte ; la lecture de divers manuels, des plus vénérables (Giry, 1894) aux plus récents (Guyotjeannin, 1993 ; Vieillard, 2001a ; Vieillard, 2001b), nous a permis de prendre conscience de ces normes déjà bien établies.

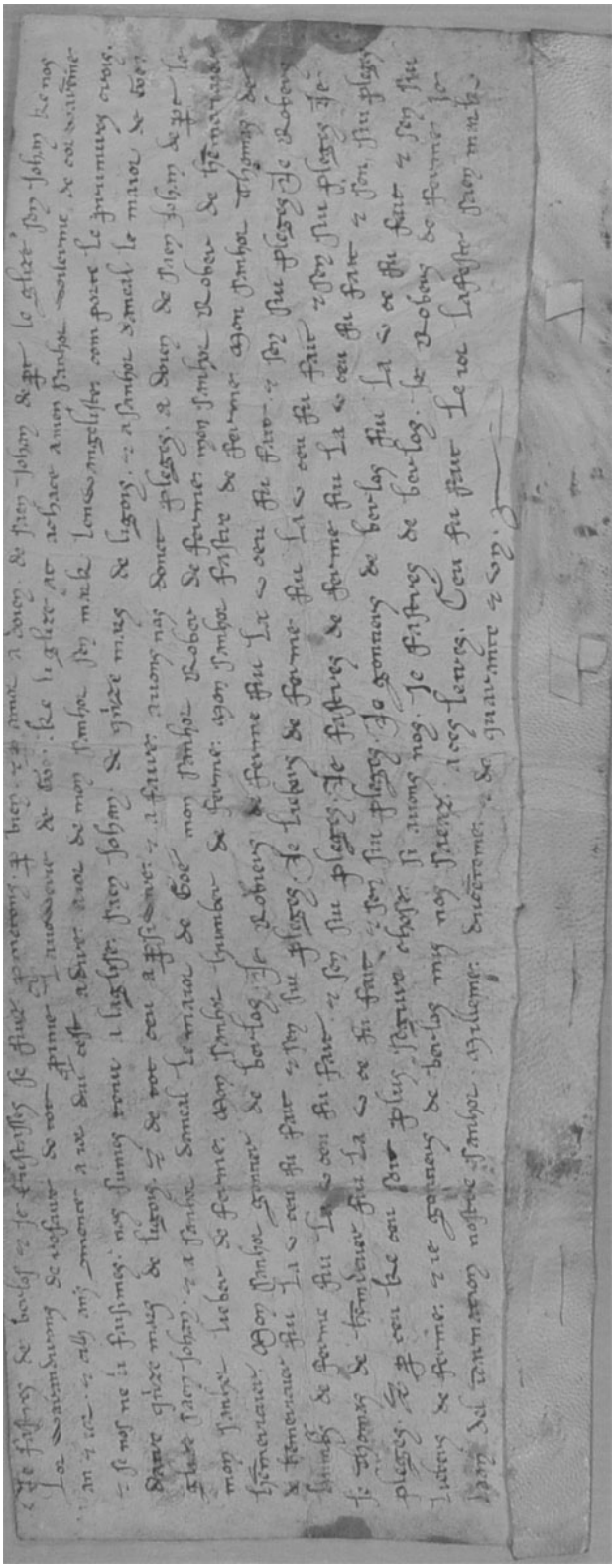
Comme nous sommes linguistes (du moins en formation), nous avons directement entrevu la possiblité de travaux linguistiques : les chartes devaient servir de matériaux pour l'étude d'unités qu'on peut définir intuitivement comme étant des *textes* (études énonciatives), des *phrases* (études syntaxiques), des *mots* (lexicologie, onomastique), des *lettres* (graphétique et graphé-

[1] Nous en profitons pour remercier Marie-Guy Boutier, qui s'est donné la peine de relire attentivement le texte de notre communication.

[2] Selon les mots de Marie-Guy Boutier.

[3] Cela n'exclut pas que l'on puisse raffiner la typologie.

Figure 1. Archives de l'État à Liège, collégiale Saint Jean l'Évangéliste, 1241



mique) et des *signes de ponctuation* (études spécifiques).

Très vite, il nous est apparu que les manuels de diplomatique, élaborés par des historiens, décrivaient des techniques insuffisantes à nos yeux de linguistes, non qu'elles soient mauvaises en soi, mais parce qu'elles réduisent de manière trop importante la richesse linguistique des documents. Nous ne trouvions guère de quoi satisfaire notre appétit dans les manuels de philologie (Foulet, 1979 ; Lepage, 2001), le plus souvent basés sur la tradition. Les conventions de ces guides d'édition permettent de donner des textes à lire à des historiens ou à des littéraires, pas à des linguistes.

Dans cette optique, un questionnement préalable sur les études linguistiques possibles sur ces textes est inévitable. De manière générale, des études de syntaxe, de lexicologie ou de dialectologie historique sont envisageables, mais aussi d'autres, considérant des aspects négligés jusqu'à présent, comme celui de la graphétique ou de la ponctuation.

C'est précisément ce dernier domaine que nous avons choisi d'étudier. Pour que l'édition puisse nous servir de base de travail, il nous fallait *préserver* ces signes jusqu'ici délaissés.

4. Le texte dans tous ses états

4.1. *Perspective traditionnelle*

Au départ, nous pensions naïvement que la question « Comment arriver à nos fins ? » recevait comme réponse « Il suffit de transcrire tout ce dont nous avons besoin ». Malheureusement, l'accumulation de tout ce que nous trouvions « intéressant » nous a vite déçus : d'une part, les documents que nous produisions étaient difficilement lisibles pour les personnes extérieures au projet, et même pour nous, lorqu'il était question de nous relire ou de nous corriger. Nous pensions de manière traditionnelle et n'entrevoyions que deux solutions. La première (a) consistait à imprimer deux versions différentes du texte, l'une en-dessous de l'autre. La seconde (b) consistait à imprimer un seul texte en s'ingéniant à le réduire le moins possible tout en y introduisant des marques d'élaboration par le truchement de marques typographiques. Par exemple, Jacques Monfrin, même s'il ne le fait pas, préconise, dès le début de la série des Documents linguistiques de la France, de conserver la ponctuation originale devant les noms propres[4]. De même, M.-D. Gleßgen marque les petites capitales originales par des grasses (idée que nous avons reprise ci-après).

Revenons au texte présenté ci-dessus (p. 795). Si nous le transcrivons en gardant tous les signes qui nous intéressent dans le cadre d'une étude sur la ponctuation, nous obtenons quelque chose comme (le signe ‹ | › indique une petite espace) :

> [1] **j**e **f**aſtres de **b**erloſ *et* je **e**uſtaſſes ſe fíuz · p*ro*metons · p*or* bıen *et* p*or* amor a doıen · de ſaen **j**ohan de | p*or*t le glıze ſen johan ke nos [2] **l**or warandırons de | reſcure de | tot p*ro*ıme **l**auowerıe de **g**oe ꝛ ke lı glıze at achate a | mon ſanhor wılerme de cor | ware*m*me · [3] an *et* ıor · *et* cıl ans *com*mence a ıor duí ceſt a | dıre · a | ıor de mon ſanhor ſen mark lenwangelıſte · com | porte le prumıres croıs · [4] *et* ſe noſ ne lı faıſımes ꝛ nos ſumes tenu a laglıſe · ſaen johan · de q*u*ınze mars de lıgoıs · *et* a | ſanhor daneal le maíor de **g**oe · [5] datre q*u*ınze mars de lıgoıs

Puisqu'il nous serait utile d'avoir accès à la ponctuation moderne également, nous aurions besoin d'une édition qui ressemblerait à :

> [1] Je, Fastrés de Berlos, *et* je, Eustasses, se fiuz, p*ro*metons, p*or* bien *et* p*or* amor a doien de Saen

[4] « Il n'y avait qu'un bien mince intérêt à côté de beaucoup d'inconvénients à tenir compte de[s] signes [de ponctuation originaux] dans la publication. En revanche, je regrette d'avoir supprimé les points (. ou ..) qui précèdent les noms propres dans quelques actes [...] ; je les conserverai à l'avenir. » (Monfrin, 1974 : lxiv)

> Johan de|*por*t le glize Sen Johan, ke nos [2] lor warandirons de rescure de tot p*ro*ime l'avowerie de Goé, ke li glize at achate a|mon sanhor Wilerme de Cor|ware*m*me, [3] an *et* jor ; *et* cil ans *com*mence a jor d'ui, c'est a|dire a|jor de mon sanhor sen Mark l'Enwangeliste c'om|porte le Prumires Crois. [4] *Et* se nos ne li faisimes, nos sumes tenu a la_glise Saen Johan de q[u]inze mars de ligois, *et* a|sanhor Daneal, le maior de Goé, [5] d'atre q*u*inze mars de ligois.

Cette démarche (a) complique le travail de l'éditeur, du correcteur, et de celui qui veut simplement comparer les deux états du texte, le poussant à effectuer sans-cesse des allers-retours entre eux (ce qui l'amène inévitablement à commettre des erreurs), c'est ce qu'on peut nommer *problème de correspondance*.

D'où l'utilité, si nous désirons comparer la ponctuation originale à la ponctuation moderne, d'une édition qui ressemblerait à :

> [1] **J**e, Faſtrés de Berloſ, *et* ȷe, **E**uſtaſſes, ſe fíuz, · p*ro*metons, · p*or* bıen *et* p*or* amor a doıen · de Saen **J**ohan dep*or*t le glıze Sen Johan, ke nos [2] **l**or warandırons de reſcure detot p*ro*ıme **l'**avowerıe de **G**oé,ˀ ke lı glıze at achate a|mon ſanhor Wılerme de Cor|ware*m*me, · [3] an *et* ȷor ; · *et* cıl ans *com*mence a ȷor d'uí, c'eſt a dıre · a|ȷor de mon ſanhor ſen Mark l'Enwangelıſte · c'om porte le Prumıres Croıs. · [4] *Et* ſe noſ ne lı faıſımes,ˀ nos ſumes tenu a la_glıſe · Saen Johan · de q[u]ınze mars de lıgoıs, · *et* a|ſanhor Daneal, le maíor de **G**oé, · [5] d'atre q*u*ınze mars de lıgoıs.

Malheureusement, cette méthode rend fastidieuse la lecture du document pour celui qui n'est pas intéressé par la ponctuation, l'encombrant de bruits, ce qui constitue un *problème de compréhension*.

Bien sûr, nous pouvions raisonner comme le fait Arrigo Castellani (1985 : 246) et privilégier la seconde solution, considérant qu'elle ne fait obstacle qu'au lecteur peu expérimenté[5]. Mais ce n'est pas exclusivement l'ignorance qui rend un texte surchargé difficile d'accès : la surcharge impose une opération de discernement supplémentaire de la part du lecteur ; et les marques de ponctuation moderne peuvent constituer une surcharge pour celui qui ne s'y intéresse absolument pas. À quoi bon fournir une édition qui ne serait pas plus facile à lire que le support original ?

En somme, la solution au problème de correspondance est le document unique et celle qui résout le problème de compréhension tout en préservant les données est le document double. D'un point de vue traditionnel, les deux sont incompatibles.

Donc, nous voici amené à faire trois éditions distinctes d'un même texte en fonction de l'usage que nous voulons en faire. On comprendra aisément que l'édition d'un simple document d'une dizaine de lignes peut rapidement mener à un casse-tête chinois lorsqu'il est question de corriger quelques erreurs de lecture ou d'apporter des précisions par des notes interprétatives. D'autre part, dans la mesure où aucune de ces trois éditions ne conviendrait à un lecteur appartenant au projet, il faudrait en produire une quatrième... La voie traditionnelle mène à l'impasse.

4.2. *Apports des nouvelles technologies*

La perspective traditionnelle ne rend possible aucun lien opératoire : dès lors qu'il en a besoin, l'humain doit toujours construire les relations qui ne figurent pas sur le papier.

[5] « Rimane la scelta tra il testo [di un trattato d'Alessandro Leonardi (1554) come è originalamente uscito] e [un] testo [...] in cui la punteggiatura, la divisione delle parole, le maiuscole e gli accenti sono conformi all'uso moderno. Darei la preferenza al primo, giacché i segni paragrafematici costituiscono un sistema sufficientemente evoluto e coerente : sistema di cui è difficile non tener conto, anche se in parte differisce da quello attuale e se può causare certe difficoltà (o certe ulteriori difficoltà, in più di quelle grafematiche) al lettore sprovveduto. Ma quanti sono i lettori sprovveduti che si dilettano a leggere il trattato di Leonardi ? E per gli altri, non basterà una semplice avvertenza sugli usi grafici e interpuntivi dell'epoca a cui risale la stampa originale ? »

Comme chacune des éditions traditionnelles n'est qu'une façon de regarder un même texte, nous avons pensé à créer un fichier contenant un document unique bardé de notes — solution traditionnelle (b) —, mais dont il serait possible de ne sélectionner qu'une partie en fonction du regard qu'on veut poser sur le document. Nous nommons cette méthode *texte monolythique*, parce qu'elle permet de maintenir l'unité intellectuelle du texte.

Imaginons que la première ligne du document présenté ci-dessus soit sauvegardée suivant un encodage fictif[6] qui marque les résolutions d'abréviation entre parenthèses (pratique par ailleurs très courante) :

```
[1] Je, Fastrés de Berlos, (et) je, Eustasses se fiuz p(ro)metons
p(or) bien (et) p(or) amor a doien de Saen Johan dep(or)t le glize
$en johan ke nos
```

Les parenthèses inélégantes sont utiles pour le sceptique lexicologue, qui veut être certain qu'un mot est bien attesté sous une forme et pas une autre ; elles lui offre un moyen de contrôle supplémentaire. Par contre, elles ne servent à rien pour l'historien. Dans cette optique, un seul fichier contient donc le texte dans tous ses états, entremêlant marques d'élaboration et richesse non réduite sous la forme d'un balisage *ad hoc*[7], mais chacun n'y sélectionne que ce qui l'intéresse. Les deux solutions traditionnelles, *a priori* antagonistes (rendu unique *vs* rendu multiple), sont conciliées.

Le texte monolithique s'impose par sa flexibilité et par la sécurité qu'il offre aux données (en rendant impossible les anomalies de mise à jour). En plus d'offrir tous les avantages combinés des méthodes traditionnelles, la méthode permet une numérotation et un classement aisé des documents.

5. Élaborer le langage

5.1. *Le choix d'XML*

Le choix de la norme XML (*eXtensible Markup Language*, v. W3C, 2000) s'est imposé assez rapidement : lisible, simple, populaire, personnalisable et extensible, mais aussi gratuit, le standard permet de créer des documents qui peuvent en outre être manipulés par une grande quantité de logiciels libres. Enfin, les fichiers associés à des feuilles de style (W3C, 1999) génèrent les formats de sortie désirés[8], ce qui répond à la nécessité de différencier le contenu du formatage.

5.2. *À la recherche de nouvelles solutions*

Du fait de la spécificité du traitement que nous réservons aux données de nos chartes, nous n'avons pas trouvé de solution satisfaisante dans le résultat du travail énorme fourni par la TEI et EAGLES pour définir des règles d'encodage flexibles utilisant XML.

La plupart du temps, les solutions les plus proches de ce que nous cherchions proposaient d'*aligner plusieurs corpus*[9]. La TEI donne des exemples où les variantes de différents manuscrits sont effectivement « alignées » à l'aide le l'apparat critique (TEI Consortium 2001 :

[6] C'est-à-dire un texte tel qu'il pourrait figurer dans le fichier informatique.

[7] Le balisage peut en effet inclure n'importe quel type d'information « Markup can be used to encode many different features within a text. In general, the more markup there is in a text, the more useful that text can be, and the more expensive it is to create. Markup can be used to encode both structural and analytic features. » (Hockey, 1998 : 108)

[8] Les formats ASCII, HTML et RTF ne posent aucun problème ; le format PDF peut être obtenu via un fichier LaTeX — solution que nous avons retenue.

[9] C'est-à-dire de mettre deux textes correspondant suivant un critère choisi en correspondance réciproque ; voir, par exemple, Habert, 1997 : 135, qui présente les corpus alignés dans la perspective de la traduction.

§19) :

> A very simple partial apparatus for the first line of the Wife of Bath's Prologue might take a form something like this :
>
> ```
> <app>
> <rdg wit="El">Experience though noon Auctoritee</rdg>
> <rdg wit="La">Experiment thogh noon Auctoritee</rdg>
> <rdg wit="Ra2">Eryment though none auctorite</rdg>
> </app>
> ```

La relation de correspondance entre les mots est évidente, mais ce système ne permettrait que difficilement de mettre un caractère en relation avec un autre ou un groupe de caractères. D'autre part, si l'alignement est tout à fait justifié pour le cas de deux manuscrits différents ou d'une traduction, sa validité est on ne peut plus douteuse dans le cas de deux vues possibles d'un texte *unique* ; sans compter qu'il mène à presque tous les désavantages de la première solution traditionnelle (document multiple) : dislocation de l'unité du texte, danger d'anomalies de mise à jour, multiplication des versions en fonction des emplois, etc.

5.3. *Une philosophie : l'économie*

Pour élaborer nos règles d'encodage, nous avons poursuivi deux objectifs : éviter ce qu'on nomme le *fatware* et privilégier la transparence des données.

5.3.1. *Éviter le* fatware

Par *fatware*, on entend « un défaut consistant, pour un logiciel ou un langage, à avoir une taille démesurée par rapport à sa fonctionnalité »[10].

D'après Gary Simons, dans le cadre d'un projet particulier, ciblé et donc nécessitant moins de ressources, l'utilisation de langages très généraux pour encoder des documents peut mener au *fatware* :

> The large SGML DTDs[11] in widespread use (e.g. HTML, DocBook, ISO 12083, CALS, EAD, TEI) offer the advantage of standardization, but for a particular project they often carry the disadvantage of being too large or too general (Simons, 1998 : §1).

Il peut arriver que la DTD TEI ne convienne pas à un projet spécifique :

> For the Sikaiana dictionary project, the TEI DTD proved to be huge in comparison to the subset of elements and attributes that were actually used (Simons, 1998 : §2.3).

D'après lui, la solution consiste à définir soi-même une DTD adéquate, ce qui est rendu aisé par le fait qu'XML est initialement prévu pour ce type de développement rapide d'un balisage *ad hoc* tout en restant conforme à un standard. En adoptant cette position, on simplifie de nombreuses tâches inévitables dans la mise en œuvre d'un projet

> Having a DTD that is limited to just the elements and attributes that are used in a project simplifies many tasks like building project-specific software, specifying stylesheets, shipping the DTD with the data, and documenting markup practice (Simons, 1998 : §2.3).

[10] « Five years ago, a cover story in Byte (Perratore, 1993) decried the problem of « fatware » — software that just keeps getting bigger and bigger with each release without returning commensurate benefit to the user. Niklaus Wirth, in his plea for lean software (Wirth, 1995), sums up the situation thus : « Software's girth has surpassed its functionality. » » (Simons, 1998 : §2.3, nous adaptons les références)

[11] [*Document Type Definition*, grammaire définissant formellement un document SGML.]

Cet écueil à éviter – tout spécialement pour un projet en développement et encodé par des linguistes peu rompus à l'informatique – constitue une des premières raisons qui nous ont menés à « contourner » les recommandations pourtant détaillées et sans cesse améliorées de la TEI (2001) ou d'EAGLES (1999). En fait, les documents que le projet présente sont d'une extrême simplicité par rapport à tout ce qui est prévu par TEI. La conversion du corpus vers ces formats généraux reste possible – notamment grâce aux transformations XSLT[12], qui permettent de convertir facilement un fichier XML en un autre. C'est ce qui nous a amené à définir un langage dédié spécialement à l'édition des chartes : KML (*Khartes Markup Language*)

5.3.2. *Privilégier la transparence*

Dans un projet linguistique à caractère philologique utilisant l'informatique, on peut s'attendre à rencontrer des problèmes récurrents liés principalement à la distance qui existe entre les deux disciplines.

Le projet *Khartês* entend intégrer des étudiants préparant un mémoire de licence (spécialistes en formation) ; et il va de soi que si l'on peut demander au linguiste de devenir philologue, il est beaucoup moins raisonnable de le forcer à être à même de comprendre et manipuler tout ce qui a trait à l'aspect informatique du projet. Néanmoins, on peut lui demander de se servir d'outils existants d'usage accessible pour le non-spécialiste.

Pour rendre l'activité d'encodage facile d'accès, il y a deux solutions : soit fournir une interface efficace à l'utilisateur, soit créer un langage et des modalités de balisage simples.

Des deux solutions, nous avons choisi la seconde, essentiellement parce que le développement d'une interface est un métier en soi et demande du temps (tant pour la conception que pour le test) et des compétences que nous n'avons pas nécessairement le loisir développer.

Nous privilégions ainsi la *transparence de l'encodage*, la capacité qu'a un encodage à signifier de lui-même ce qu'il enregistre, c'est-à-dire à être lisible par un être humain – n'importe quel encodage est lisible par un ordinateur, pourvu qu'on lui fournisse le bon algorithme. Pour un humain, `u0041`, `A` et `A` pour signifier *A* sont illisibles, et personne ne reconnaîtra *livre* dans

```
&#x006C;&#x0069;&#x0076;&#x0072;&#x0065;
```

Mais si la solution idéale est de trouver un signe motivé pour tout élément encodé, sa mise en pratique est difficile. Nous devons en effet nous plier à des contraintes purement matérielles (et contingentes), comme le nombre de touches d'un clavier par exemple. Pour la plupart des caractères, cela va sans difficulté, mais lorsqu'il s'agit d'encoder une analyse et que nous avons besoin du balisage, il faut que ce dernier soit bref et explicite ; et ces deux attributs, applicables tant au balisage qu'aux données caractères, sont souvent antagonistes...

L'inconvénient est qu'en évitant les entités Unicode[13], les données ne sont pas encodées à l'aide d'un langage strictement conforme à XML, mais une simple conversion à l'aide d'expressions rationnelles permettra de résoudre le problème lorsque le corpus devra être distibué en dehors des membres du projet.

Il est parfois nécessaire d'encoder des caractères qui n'ont pas d'équivalent Unicode défini. C'est le cas des deux variantes du signe de ponctuation nommé (Parkes, 1992) *punctus elevatus*

[12] C'est-à-dire *eXtensible Stylesheet Language Transformations*, un langage de programmation de très haut niveau destiné à modifier la structure des documents XML ; XSLT est conforme au standard XML (voir W3C, 1999).

[13] Voir http ://www.unicode.org/.

(point renforcé d'un trait ou d'un crochet) : ‹ ? › et ‹ ! ›. Puisque ces signes ressemblent fortement à ‹ ? › et ‹ ! ›, ce sont ces saisies clavier qui servent à les encoder. De la même façon, ‹ * › sert à noter ‹ · › et ‹ / › équivaut à ‹ / ›.

6. Exemple

6.1. *Document*

Nous ne pourrions faire ici une présentation complète de la DTD de KML et des conventions d'encodage pour les caractères spéciaux. On retiendra essentiellement :

– que l'élément c entoure les lettres marquées sur le support,
– que l'élément l marque le début des lignes,
– que les crochets entourent les abréviations résolues,
– que le signe ‹ ˚ › précède les caractères qui doivent être transformés (‹ u › en ‹ v › et vice-versa, ‹ | › en ‹ - ›, etc.) pour qu'un lecteur moderne appréhende plus facilement le texte.

Si nous transcrivons intégralement le document présenté p. 795, nous obtenons la transcription qui suit :

```
<charte fonds="LgSJe" an="1241(29b)"><texte><recto><part /><l /><c>J</c>e, Fa$trés de Berlo$ [et] je, <c>E</c>u$ta$$es, $e fiuz, *
<part /> p[ro]metons, * p[or] bien [et] p[or] amor a doien * de §aen <c>J</c>ohan de˚|p[or]t le glize §en Johan, ke nos <l /><c>l</c>or
warandirons de|re$cure de|tot p[ro]ime <c>l</c>'a˚uowerie de <c>G</c>oé,? ke li glize at achate a|mon $anhor Wilerme de Cor|ware[m]me,
* <l />an [et] ˚ior; * [et] cil ans [com]mence a ˚ior d'uí, c'e$t a|dire * a|˚ior de mon $anhor $en Mark l'Enwangeli$te * c'om|porte
le Prumires Crois. * <l />[Et] $e no$ ne li fai$imes,? nos $umes tenu a la_gli$e * §aen Johan * de q[u]inze mars de ligois, * [et]
a|$anhor Daneal, le maior de <c>G</c>oé, * <l />d'atre q[u]inze mars de ligois. * [Et] de tot ce a p[or]$iwre * [et] a faire * a˚uons
nos donet pleges * a doien de §aen <c>J</c>ohan de-p[or]t le <l />glize §aen Johan * [et] a $anhor Daneal le maior de <c>G</c>oé,
mon $anhor <c>R</c>ober de|Ferme, * mon $anhor <c>R</c>ober de He[m]mericuer, <l />mon $anhor Lieber de Ferme, * <c>m</c>on $anhor
Humber de Ferme, * <c>m</c>on $anghor Fa$tré de Ferme, <c>m</c>on $anhor <c>T</c>homas de <l />He[m]mericuer, * <c>m</c>on $anhor
Gontier de Berlos. * [Et]_<c>je</c><rem>l'abréviation rentre dans le <c>j</c>e</rem>, <c>R</c>obiers de Ferme, fui <c>l</c>a ˚v ceu
fu fait; [et]|$'en $ui pleges. * [Et]_<c>j</c>e, <c>R</c>ober$ <l />de He[m]mericuer, fui <c>l</c>a ˚v ceu fu fait; * [et]|$'en
$ui pleges. * [Et]_<c>je</c>, <c>L</c>iebers de Ferme, fui <c>l</c>a ˚v ceu fu fait; [et] $'en $ui pleges. * [Et]_<c>j</c>e, <l
/>Humb[er]s de Ferme, fui <c>l</c>a ˚v ceu fu fait; * [et] $'en $ui pleges. * [Et]_<c>je</c>, Fa$trés de Ferme, fui <c>l</c>a ˚v ceu
fu fait; [et] $'en $ui pleges. <l /><c>J</c>e, <c>T</c>homas de He[m]m[er]icuer, fui <c>l</c>a ˚v ce fu fait; [et] $'en $ui pleges.
* [Et]_<c>je</c>, Gontiers de Berlos, fuí <c>l</c>a ˚v ce fu fait; [et] $'en $uí <l />pleges. * <part /> [<c>Et</c>] p[or] ceu ke ceu
$oit plus $egure cho$e, $i a˚uons nos, * <c>j</c>e, Fa$trés de Berlos, * <c>j</c>e, <c>R</c>obers de Ferme, * <c>j</c>e, <l />Liebers
de Ferme, * [et] ˚ie, Gontiers de Berlos, mis nos $aeaz * a|ces letres. * <c>C</c>eu fu fait <c>l</c>e ˚ior <c>l</c>a_fe$te $aen Mark,
* <l />l'aan del I[n]carnation No$tre §anhor * <c>m</c>illeme * duce[n]teme * [et] de quarante [et] ˚vn. * <rem>signé à droite</rem>
</recto></texte></charte>
```

6.2. *Exemple d'application*

Prenons un problème concret : nous voulons étudier les lettres marquées[14] apparaissant à l'initiale des noms propres sur le support original. L'encodage d'un texte monolythique permet d'interroger ce dernier de manière à en extraire tout ce qui est formellement défini. Comme on l'a vu ci-dessus, les lettres marquées sont transcrites dans l'élément c et les capitales modernes figurent en majuscules dans le code, ce qui constitue deux marquages compatibles.

En écrivant un simple programme d'extraction, il est aisé de relever tous les mots commençant par une majuscule et de les pointer par un point d'exclamation s'ils comportent une initiale marquée. Nous obtenons la liste suivante :

1. ! Je
2. Fastrés
3. Berlos
4. ! Eustasses
5. Saen
6. ! Johan
7. Sen
8. Johan
9. ! Goé
10. Wilerme
11. Cor|waremme
12. Mark
13. Enwangeliste
14. Prumires
15. Crois
16. Et
17. Saen
18. Johan
19. Daneal
20. ! Goé
21. Et
22. Saen
23. ! Johan
24. Saen
25. Johan
26. Daneal
27. ! Goé
28. ! Rober
29. Ferme
30. ! Rober
31. Hemmericuer
32. Lieber
33. Ferme
34. Humber
35. Ferme
36. Fastré
37. Ferme
38. ! Thomas
39. Hemmericuer
40. Gontier
41. Berlos
42. Et
43. ! Robiers
44. Ferme
45. Et
46. ! Robers
47. Hemmericuer
48. Et
49. ! Liebers
50. Ferme
51. Et
52. Humbers
53. Ferme
54. Et
55. Fastrés
56. Ferme
57. ! Je
58. ! Thomas
59. Hemmericuer
60. Et
61. Gontiers
62. Berlos
63. ! Et
64. Fastrés
65. Berlos
66. ! Robers
67. Ferme
68. Liebers
69. Ferme
70. Gontiers
71. Berlos
72. ! Ceu
73. Mark
74. Incarnation
75. Nostre
76. Sanhor

[14] Lettres se distingant de leurs voisines par leur épaisseur, leur taille ou leur couleur ; elles font intuitivement penser à nos majuscules modernes.

Sur 76 mots comportant une majuscule dans l'édition, 18 mots ont été pointés comme commençant par une lettre marquée dans l'original. Parmis ces 18 mots, trois ne sont pas des noms propres. Donc, la plupart du temps, le scribe ne marque pas les noms de personnes.

Dès lors, l'emploi des lettres marquées est-il lié à la segmentation en phrases ? Un dépouillement similaire montre qu'à part la première phrase du texte, qui commence par *Je*, toutes les phrases commencent par *Et* et une seule par *Ceu*. Les lettres marquées au début de

1. ! Eustasses	4. ! Goé	7. ! Rober	10. ! Robiers	13. ! Je
2. ! Johan	5. ! Johan	8. ! Rober	11. ! Robers	14. ! Thomas
3. ! Goé	6. ! Goé	9. ! Thomas	12. ! Liebers	15. ! Robers

ne correspondent pas à des débuts de phrase. Par contre, ce qui frappe dans cette liste, c'est que les lettres qui commencent ces mots appartiennent à un stock relativement réduit : ‹ e ›, ‹ g ›, ‹ j ›, ‹ l ›, ‹ r ›, et ‹ t › ; un autre dépouillement du même texte montre que ‹ e ›, ‹ r › et ‹ t › n'apparaissent jamais non marquées au début d'un nom propre. Nous décelons donc que la tendance d'un nom propre à être graphié avec une initiale marquée dépend de la nature de cette initiale, fait qui mériterait d'être étudié sur base de textes plus nombreux de la même époque.

7. Conclusion

Avec le développement de la linguistique de corpus et la démocratisation des moyens informatiques, l'édition de texte prend une dimension nouvelle : au lieu de se limiter à « donner un texte à lire » à une catégorie restreinte de chercheurs (voir à un seul lecteur), l'éditeur peut destiner son travail à un plus vaste public, libre de réduire comme il l'entend ce qui ne lui est pas utile. La réduction de la richesse originale du document n'affecte que la *vue* désirée par le lecteur et n'altère en rien le fichier contenant l'édition. Les possibilités de comparaison entre la forme originale et le marquage moderne peuvent être partiellement (sinon totalement) automatisées, ce qui facilite grandement les dépouillements.

Références

Castellani A. (1985). Problema di lingua, di grafia, di interpunzione nell'allestimento dell'edizione critica. In Malato (1985) : 229-54.

EAGLES [= Expert Advisory Group on Language Engineering Standards] (1999). *Corpus Encoding Standard. Version 1.5*. [http ://www.ilc.pi.cnr.it/EAGLES/home.html].

Foulet A. et Speer M.B. (1979). *On editing Old French Texts*. Kansas : The Regents Press of Kansas.

Gigot J. G. (1974). *Chartes en langue française antérieures 1271 conservées dans le département de la Haute-Marne*. Paris : C.N.R.S. (Documents linguistiques de la France, série française 1).

Giry A. (1894). *Manuel de diplomatique*. Paris : Hachette.

Guyotjeannin O., Picke J. et Tock B. M. (1993). *Diplomatique médiévale*. Louvain : Brepols (L'atelier du médiéviste 2).

Habert B., Nazarenko A. et Salem A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin (U Linguistique).

Hockey S. (1998). Textual Databases. *Lawler* : 101-37.

Lawler J. et Aristar Dry H. (dir.) (1998). *Using Computers in Linguistics. A Practical Guide*. London/New York : Routledge.

Lepage Y. G. (2001). *Guide de l'édition de textes en ancien français*. Paris : Champion.

Malato E. (dir.) (1985). *La critica del testo. Problemi di metodo ed esperienze di lavoro. Atti del convegno di Lecce 22–26 ottobre 1984*. Roma : Salerno Editrice.

Mantou R. (1987). *Chartes en langue française antérieures à 1271 conservées en Flandre Orientale et Flandre Occidentale*. Paris : C.N.R.S. (Documents linguistiques de la Belgique romane 2).

Monfrin J. (1974). Introduction. In Gigot (1974) : xi-lxxx.

Parkes M.B. (1992). *Pause and effect. An introduction to the history of punctuation in the West*. Cambridge : Scholar Press.

Perratore E., Thompson T., Udell J. et Malloy E. (1993). Fighting fatware. *Byte* : 98-108.

Ruelle P. (1984). *Chartes en langue française antérieures à 1271 conservées dans la province de Hainaut*. Paris : C.N.R.S. (Documents linguistiques de la Belgique romane 1).

Simons G. F. (1998). Using architectural processing to derive small, problem-specific XML applications from large, widely-used SGML applications. *SIL Electronic Working Papers*. [http ://www.sil.org/silewp/1998/006/silewp1998-006.html].

TEI Consortium. (2001). TEI P4. Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition. [http ://www.tei-c.org].

Vieillard F. et Guyotjeannin O. (dir.) (2001a). *Conseils pour l'édition de textes médiévaux. Fascicule I : Conseils généraux*. Paris : CTHS École des Chartes.

Vieillard F. et Guyotjeannin O. (dir.) (2001b). *Conseils pour l'édition de textes médiévaux. Fascicule II : Actes et documents d'archives*. Paris : CTHS École des Chartes.

W3C [= World Wide Web Consortium]. (2000, 1998). Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommandation 6 October 2000. [http ://www.w3.org/TR/REC-xml].

W3C [= World Wide Web Consortium]. (1999). XSL Transformations (XSLT) 1.0. W3C Recommandation 16 November 1999. [http ://www.w3.org/TR/1999/REC-xslt].

Wirth N. (1995). A plea for lean software. *IEEE Computer* : 64-8.

Secondo gli esperti: Popolazione e società nelle opinioni dei testimoni privilegiati di una ricerca sugli adolescenti italiani

Maura Misiti[1], Simona Carbone[2]

[1]IRPPS – Istituto di Ricerca sulla Popolazione e le Politiche Sociali – CNR – 128, via Nizza 00198 Roma – Italy – m.misiti@irpps.cnr.it

[2]IRMC – Istituto di Ricerca sulla Multimedialità e la Comunicazione – Provincia di Caserta – Via Santa Chiara, pal. Mirabella – 81100 Caserta – Italy – simona.carbone@tin.it

Abstract

Data used in this work have been collected in three surveys carried out by IRPPS in some Italian cities. Surveys have the main purpose to collect attitudes and opinions of Italian adolescents (aged 13-14 and 16-18) towards recent population issues such as family trends, migration and gender roles. In the research both quantitative and qualitative methods are used. The quantitative survey on student's attitudes in fact, follows a number of face-to-face interviews to key-persons, selected as experts in the field of education (parents, teachers, school managers, local administrators and leaders of associations related to youth issues). By applying lexical analysis techniques on the texts of the interviews conducted during the field work in the cities participating to the survey (Milan, Salerno, Venezia) text contents are synthesised without any coding. The purpose is to grasp the structure and the patterns of the contents displayed by the language, to compare views and opinions towards the different topics of the survey among the cities. Further results are expected by the lexical correspondences analysis: this is to outline the overall structure of the content, considering differences and similarities in the language used by the interviewed key persons in the three cities. The analysis includes: frequency analysis on single forms and repeated segments, specific and banal forms, and modal phrases.

Riassunto

I dati utilizzati in questo lavoro provengono da una serie di indagini condotte dall'IRPPS in diverse città italiane finalizzate a raccogliere atteggiamenti ed opinioni degli adolescenti (14 e 16-18 anni) verso le trasformazioni demografiche recenti. Il progetto prevede, oltre alla somministrazione di un questionario agli studenti delle scuole medie e superiori, una rilevazione qualitativa sugli stessi temi, effettuata presso "testimoni privilegiati" attraverso interviste in profondità. Questi soggetti selezionati ad hoc rivestono un ruolo chiave nel settore indagato, sono esperti in quel settore, quindi le opinioni e le valutazioni da loro espresse sono molto rilevanti Gli intervistati sono genitori di adolescenti, insegnanti, dirigenti scolastici, funzionari di amministrazioni locali e presidenti di associazioni operanti nel settore degli interventi in favore dei giovani. Mediante l'analisi lessicale del testo delle interviste si sono ricercate le strutture di contenuto, per ricostruire le opinioni degli intervistati in una chiave di lettura trasversale, ovvero confrontando le tre città rispetto agli argomenti emergenti per ogni tematica affrontata. Inoltre, con l'analisi delle corrispondenze lessicali, si è ricercata la struttura complessiva del contenuto, cogliendo associazioni e contrapposizioni tra le forme usate dagli intervistati, distinguendo per città di riferimento e area tematica del questionario L'analisi lessicale è stata condotta sulle singole forme e sui segmenti ripetuti, inoltre si è approfondito il livello di ricerca dei contenuti, attraverso l'analisi delle forme specifiche utilizzate con maggiore frequenza, distinguendo per città di appartenenza, in modo da cogliere le differenze di punti di vista e di realtà sociali a livello territoriale rispetto agli argomenti in esame.

Parole chiave: tendenze demografiche, testimoni privilegiati, atteggiamenti degli adolescenti, analisi lessicale, forme specifiche, corrispondenze lessicali.

1. Introduzione

Questa ricerca costituisce la seconda fase di un'indagine condotta dall'IRPPS per rilevare gli atteggiamenti degli adolescenti verso le recenti dinamiche demografiche, migratorie e familiari. A completamento delle informazioni, quantitative, raccolte somministrando un questionario ad un campione di studenti dell'ultimo anno delle scuole superiori e delle scuole medie, la fase dell'indagine di cui in questa sede sono riportati i risultati è consistita in una rilevazione qualitativa sugli stessi temi, effettuata presso "testimoni privilegiati". Si tratta di soggetti che rivestono un ruolo chiave nel settore indagato, sono esperti in quel settore, quindi le opinioni e le valutazioni da loro espresse sui fenomeni oggetto di interesse sono molto rilevanti. Gli intervistati per questa analisi sono genitori di adolescenti, insegnanti, dirigenti scolastici, funzionari di amministrazioni locali e presidenti di associazioni operanti nel settore degli interventi in favore dei giovani. L'interesse è verso la conoscenza di questi soggetti dei fenomeni sociali ambito dell'indagine, conoscenza determinata anche dalla loro professione.

2. Metodologia della ricerca

Le interviste ai testimoni privilegiati hanno lo scopo di delineare il contesto dell'indagine, individuando il punto di vista degli attori istituzionali e locali sulle tematiche principali. Lo strumento utilizzato è quello delle interviste in profondità a protagonisti del mondo della scuola (presidi ed insegnanti), dell'Ente locale (funzionari nel settore delle Politiche Sociali, delle Politiche Giovanili, delle Politiche Educative) e della realtà territoriale. In totale sono stati intervistati 29 soggetti; il criterio di selezione tiene conto della rappresentanza in base al genere, anche se si sono trovati più uomini al vertice di amministrazioni e associazioni, mentre tra gli insegnanti vi è una netta prevalenza femminile. In ogni caso, si è ritenuto che le opinioni espresse fossero principalmente determinate dalla loro competenza settoriale, e solo in parte influenzate dall'appartenenza di genere. Gli intervistati selezionati vivono e lavorano in tre città, Milano, Venezia e Salerno, e pertanto sono competenti ciascuno della realtà sociale, politica e culturale della propria città, oltre che del Paese nel complesso.

Le interviste sono state condotte da un intervistatore seguendo una traccia, articolata in quattro grandi sezioni tematiche che ricalcano la struttura del questionario proposto agli studenti, sulla base di questa struttura comune le interviste agli insegnanti/presidi prevedono questioni specifiche relative alle eventuali ricadute didattiche del progetto, mentre quelle agli amministratori, genitori, rappresentanti dell'associazionismo sono domande relative alla pianificazione, implementazione e gestione di interventi concernenti le tematiche dell'indagine. La parte comune prevede le seguenti tematiche: *Recenti tendenze demografiche nel mondo, in Europa, in Italia, nella Regione e nel comune di appartenenza; immigrazione in Italia, nella Regione e nel comune di appartenenza; nuove tipologie di famiglia; ruoli di genere*. Agli intervistati operanti nel comune di Milano è stata proposta anche una quinta sezione, relativa alla rilevazione degli atteggiamenti verso il futuro, il lavoro e la società. Per ogni sezione, agli intervistati è stata chiesta un'opinione circa la conoscenza e l'interesse dei giovani sull' argomento, l'opportunità che questo venga trattato ed approfondito nelle scuole, inserito nei programmi scolastici, e di riferire sulle azioni e gli interventi di loro conoscenza realizzati nel comune per risolvere eventuali problemi connessi e divulgare l'informazione sulla tematica. Mediante l'analisi lessicale del testo delle interviste si sono ricercate le strutture di contenuto, per ricostruire le opinioni degli intervistati in una chiave di lettura trasversale, ovvero confrontando le tre città rispetto agli argomenti emergenti per ogni tematica affrontata. Inoltre, con l'analisi delle corrispondenze lessicali, si è ricercata la struttura complessiva del contenuto, cogliendo associazioni e contrapposizioni tra le forme usate dagli intervistati, distinguendo per città di riferimento e area tematica del questionario. La sezione relativa agli

atteggiamenti verso il futuro, il lavoro e la società, presente solo per gli intervistati di Milano, è stata esclusa per questa parte dell'analisi, per omogeneità di contenuto tra le città. L'utilizzo di una traccia strutturata per l'intervista, ma soprattutto la specificità degli argomenti trattati, che ha lasciato poco spazio a divagazioni e che richiede, per sua natura, una terminologia settoriale specialistica, hanno determinato una scarsa ambiguità nel vocabolario del testo, in cui si ritrovano spesso anche termini utilizzati per la formulazione delle domande. Ad esempio, la parola "classi" viene utilizzata sempre in ambito scolastico, e solo due volte in riferimento alle "classi sociali"; una volta individuata la sequenza "in genere", la forma "genere" si riferisce sempre alle differenze in base al sesso di appartenenza. Non si è ritenuto necessario, pertanto, agire in maniera eccessiva sul testo, se non fondendo, solo per la prima parte dell'analisi, le frequenze di alcune forme, ritenute importanti perché rappresentative di temi di base, presenti nelle varie flessioni, per ottenere il massimo potere informativo. Come riferito dagli intervistatori, l'atteggiamento degli intervistati è stato mediamente favorevole, salvo l'imbarazzo creato dall'affrontare alcuni temi (quale quello relativo alle tendenze demografiche), ritenuti di poca importanza o poco conosciuti dal soggetto, che aveva comunque l'intenzione di mostrarsi competente, e pertanto è stato infastidito dall'essere colto impreparato. Inoltre, l'intera intervista ha richiesto, in media, poco meno di 2 ore per essere completata, ed alcuni intervistati, dopo essersi dilungati sulle risposte alle prime domande, hanno poi mostrato una certa insofferenza per la durata dell'impegno, riducendo progressivamente i tempi, e dunque l'approfondimento, delle risposte alle domande delle ultime sezioni dell'intervista.

Sub- testi	Occorrenze	Forme grafiche	Hapax	$G = \frac{V}{\sqrt{N}}$
		MILANO		
Demografia	27.598	4.487	2.515	10,81
Immigrazione	19.039	3.530	2.089	8,50
Nuove famiglie	11.257	2.404	1.414	5,79
Ruoli di genere	5.651	1.497	927	3,61
Futuro, lavoro, società	12.528	2.567	1.538	6,18
Totale Milano	***76.073***	***8.268***	***4.300***	***19,91***
		VENEZIA		
Demografia	13.862	2.869	1.661	6,91
Immigrazione	28.292	4.359	2.397	10,50
Nuove famiglie	16.632	2.925	1.638	7,04
Ruoli di genere	7.583	1.774	1.033	4,27
Totale Venezia	***66.369***	***7.354***	***3.806***	***17,71***
		SALERNO		
Demografia	4.798	1.411	870	3,40
Immigrazione	12.573	2.763	1.613	6,65
Nuove famiglie	9.408	2.200	1.325	5,30
Ruoli di genere	3.199	1.032	668	2,49
Totale Salerno	***29.978***	***4.761***	***2.639***	***11,47***
		SEZIONI DELL'INTERVISTA		
Demografia	46.258	6.263	3.392	15,08
Immigrazione	59.904	7.270	3.866	17,51
Nuove famiglie	36.607	5.062	2.741	12,19
Ruoli di genere	17.123	3.139	1.823	7,56
Totale Corpus	**172.420**	**13.181**	**6.322**	**V/N=7,64**

Tabella 1. Caratteristiche del corpus

3. Caratteristiche del corpus

Il testo analizzato è composto da 171.420 occorrenze, con 13.181 forme grafiche, 6.322 hapax, le parole con frequenza pari ad 1. La ricchezza lessicale, ovvero la proporzione di parole diverse sul totale delle occorrenze, è pari al 7,6% (Tabella 1). Distinguendo per città,

gli intervistati di Milano hanno dimostrato una maggiore ricchezza lessicale (19,9%), seguiti da quelli di Venezia (17,7%). Il linguaggio meno ricco è quello degli intervistati di Salerno (11,5%). La ricchezza lessicale per i sub-testi è stata calcolata mediante il coefficiente *G di Giraud* (Giraud, 1960), che tiene conto della differenza di ampiezza dei sub-testi. In effetti, il sub-testo delle interviste realizzate a Salerno risulta essere il più piccolo, con 29.978 occorrenze, quello di Venezia il più grande (66.369 occorrenze), mentre quello di Milano è intermedio, 76.073 occorrenze, che diventano 63.545 escludendo il testo delle risposte alla sezione 5 del questionario. L'argomento che presenta il maggior numero di occorrenze, è l'immigrazione (59.904 occorrenze), seguito dagli aspetti demografici della popolazione (46.258 occorrenze). A seguire le altre sezioni, con frequenze decrescenti. In realtà gli argomenti che mostrano frequenze maggiori sono quelli trattati nelle prime due sezioni del questionario; la diminuzione delle frequenze per gli altri argomenti si può spiegare con la lunghezza dell' intervista: quando hanno affrontato le ultime sezioni del questionario gli intervistati erano stanchi e desiderosi di concludere, quindi sbrigativi nelle risposte. Mentre per Salerno e Venezia il numero delle occorrenze maggiore è per le risposte sul tema dell'immigrazione, per Milano è sulla demografia.

FORMA	FREQ.	FORMA	FREQ.	FORMA	FREQ.
scuol#	957	religion#	121	comportament#	53
ragazz#	803	difficoltà	118	responsabilità	53
problem#	626	immigrazione	110	associazion#	52
famigli#	559	integrazione	106	inserimento	51
credo	509	lavorare	103	ambiente	50
figl#	491	esperienz#	102	omosessual#	48
lavoro	430	diritt#	100	comunità	47
donn#	290	popolazion#	100	dinamic#	42
cultur#	275	insegnar#	94	clandestin#	38
penso	258	sesso#	91	educator#	37
insegnant#	233	tematic#	90	volontariato	32
immigrat#	215	sposar#	88	opportunità	31
social#	211	formazione	86	disagio	30
class#	193	matrimon#	86	guerr#	30
important#	192	differenz#	84	educativo	29
lingu#	189	problematic#	82	mercato	29
educazione	186	storia	78	television#	29
giovan#	180	conoscenza	77	diversità	28
stranier#	171	diverso	77	invecchiamento	27
rapport#	168	intervent#	74	multicultur#	26
mondo	160	demografi#	70	adolescenza	24
politic#	158	femmin#	68	mentalità	24
società	156	strument#	68	profughi	21
argoment#	152	lavor#	65	etni#	20
coppia	149	arab#	64	solidarietà	20
gener#	149	valori	64	percezione	19
serviz#	148	generazion#	61	trasformazion#	18
discorso	147	intercultural#	61	parità	17
progett#	142	adolescent#	59	convivenza	16
fenomen#	135	cattolic#	57	regolar#	14
attività	134	accoglienza	54	extracomunitari	12
tem#	133	insegnamento	54	soggiorno	12
direi	124	mediator#	54	permesso	11

Tabella 2. Selezione di forme del vocabolario per frequenza

4. Analisi lessicale

L'analisi lessicale è stata condotta sulle singole forme e sui segmenti ripetuti. Le parole più frequenti sono risultate essere quelle strumentali alla costruzione delle frasi (che, di, e,...). Selezionando le parole di contenuto, tra quelle con le frequenze più alte si trovano gli argomenti di interesse (Tabella 2). Si osserva la presenza delle forme "credo", "penso", "direi", che indicano l'espressione di una opinione. In particolare, si segnala l'uso delle forme "problem#", "problematic#", "tem#", "tematic#", "fenomen#", "argoment#": le forme con radice 'problem' risultano avere, in totale, frequenza di poco inferiore al totale di tutte le altre forme, sinonimi per definire l'oggetto della discussione. Bisogna specificare che nella traccia sono usate le stesse parole, ma la forma "problem#" è usata solo in riferimento a situazioni di difficoltà relative ai temi.

SEGMENTI	FREQ.	SEGMENTI	FREQ.	SEGMENTI	FREQ.
secondo me	307	mercato del lavoro	20	campi profughi	9
non so	192	mediatori_culturali	17	lingua araba	9
penso che	154	non penso	17	permesso di soggiorno	9
come dire	149	non saprei	17	pari opportunità	9
io credo	108	ragazzi_stranieri	16	rapporti di genere	9
non lo so	82	mondo del lavoro	16	dal mio punto di vista	9
io penso	75	non mi sembra	14	pubblica amministrazione	9
mi pare	74	meno figl#	13	all'interno della scuola	9
non credo	74	alunni_stranieri	13	dinamiche demografiche	8
io sono	65	senso di responsabilità	13	asili nido	8
voglio dire	65	educazione alla sessualità	25	coppie di fatto	8
mi sembra	63	tra di loro	12	condizioni di vita	8
che vengono	55	posto di lavoro	11	bambini stranieri	7
direi che	55	politiche_sociali	11	al di fuori del matrimon#	7
gli altri	49	programmi scolastici	11	nucleo familiare	7
queste persone	46	comportamenti a rischio	10	tempo prolungato	7
devo dire	46	fare figl#	10	possibilità di lavoro	7
si parla	40	la mia opinione	10	andare a scuola	7
questi ragazzi	40	che ne so	10	calo demografico	6
per loro	34	punto di riferimento	10	storia delle religioni	6
coppi# omosessual#	32	ruolo della scuola	10	famiglie monogenitoriali	6
religione_cattolica	25	associazioni_di_volontariato	11	vita matrimon#	5

Tabella 3. Selezione di segmenti ripetuti per frequenza

In sostanza, dalla frequenza di uso di queste forme, oltre all'effetto dell'economia di linguaggio, che porta a ripetere nella risposta termini appena ascoltati nella domanda, emerge, in prima analisi, una percezione degli argomenti in esame come 'problemi', quindi in accezione negativa, derivante probabilmente da stereotipi culturali assimilati. Anche dall'osservazione dei segmenti ripetuti più frequenti emergono i temi in discussione (Tabella 3). Numerosissime sono le forme che, nel discorso, indicano l'espressione di una opinione o un punto di vista. Tra i più frequenti, "io credo", "come dire", "io penso", "direi che". Molto frequenti sono anche le sequenze "non lo so", "non so", "non mi pare", che denotano scarsa conoscenza dell'argomento trattato. Interessante è anche la presenza di "gli altri" e "queste persone", modi di definire soggetti da cui si prendono le distanze, come anche "per loro" e "tra di loro".

5. I contenuti specifici: un confronto geografico

Dall'analisi del vocabolario si sono ottenute delle informazioni generali sui contenuti; in questa fase sono state ricercate le forme usate dagli intervistati distinguendo per città di

appartenenza, in modo da cogliere le differenze di punti di vista e di realtà sociali. Si sono selezionate, pertanto, le forme specifiche utilizzate con maggiore frequenza, distinguendo per città. Le forme banali, ovvero quelle presenti in proporzioni simili nei sub-testi, non sono state prese in esame, in quanto sono risultate essere parole grammaticali e le parole che indicano in modo generico i temi dell'intervista.

MILANO				VENEZIA				SALERNO			
Forme testuali	**F**	**f**	**Spec.**	**Forme testuali**	**F**	**f**	**Spec.**	**Forme testuali**	**F**	**f**	**Spec.**
FORME TESTUALI SPECIFICHE											
insegnanti	141	97	+E09	servizi	89	80	+E24	società	156	87	+E27
educazione alla sessualità	25	25	+E09*	secondo me	307	183	+E12	religione	92	51	+E16
progetto	85	65	+E09	problema	323	179	+E10	penso che	158	72	+E16
loro	663	363	+E08	profughi	21	21	+E09*	lavoro	430	123	+E14
sono temi	18	17	+E06	mi pare	74	55	+E08	immigrati	178	70	+E12
argomenti	70	47	+E05	bambini	157	91	+E07	tematiche	74	39	+E12
soldi	41	31	+E05	minori	26	22	+E06	ritengo che	32	22	+E10
minoranza	14	14	+E05*	rapporto	104	63	+E06	problematiche	67	32	+E08
interculturale	14	14	+E05*	non so	192	100	+E05	credo che	284	87	+E08
politica	64	44	+E05	centro donna	11	11	+E05*	immigrazione	110	41	+E07
esperti	21	17	+E04	atteggiamento	59	38	+E05	senso di responsabilità	12	10	+E06
mercato del lavoro	20	17	+E04	coppia omosessuale	11	10	+E04	invecchiamento	27	14	+E05
internet	14	13	+E04	mediatori culturali	17	14	+E04	regolare	11	8	+E05
adesso non so	10	10	+E04*	anziani	36	24	+E04	problemi	279	71	+E04
comportamenti a rischio	10	10	+E04*	associazioni di volontariato	11	10	+E04	coppie omosessuali	21	8	+E04
computer	12	10	+E03	queste persone	46	29	+E03	scuola	652	151	+E04
alfabetizzazione	14	11	+E03	il diverso	11	9	+E03	insegnare	33	14	+E04
ragazzi stranieri	16	13	+E03	vita quotidiana	10	9	+E03	mentalità	24	10	+E03
multiculturalità	10	9	+E03	rifugiati	10	9	+E03	tradizionale	10	6	+E03
argomento	82	47	+E02	clandestini	16	12	+E03	garantire	21	10	+E03
alunni stranieri	13	10	+E02	matrimonio	61	34	+E03	nei programmi	20	8	+E02
solidarietà	20	13	+E02	politiche sociali	11	9	+E02	pari opportunità	9	5	+E02
FORME TESTUALI RARE											
rifugiati	10	1	-E02	scuola	652	224	-E02	mercato del lavoro	20	0	-E02
mediatori culturali	17	3	-E02	ragazzi stranieri	16	2	-E02	mediatori	17	0	-E02
stereotipo	14	2	-E02	educazione dei figli	11	1	-E02	sostegno	21	0	-E02
io ritengo che	10	0	-E03	meno figli	12	1	-E02	stranieri	120	9	-E03
associazioni di volontariato	11	0	-E03	mercato del lavoro	20	3	-E02	educazione alla sessualità	25	0	-E03
clandestini	16	1	-E03	inseriti	21	4	-E02	insegnanti	141	9	-E04
accoglienza	54	11	-E04	interculturale	14	0	-E03	soldi	41	0	-E04
coppie omosessuali	21	0	-E04	minoranza	14	0	-E03	adolescenti	45	0	-E04
profughi	21	0	-E06	integrazione	106	28	-E03	progetto	85	3	-E05
invecchiamento	27	0	-E07	senso di responsabilità	12	0	-E03	servizi	89	2	-E06
problema	323	71	-E17	insegnanti	141	35	-E04	non so	192	12	-E06
religione	92	3	-E19	tematiche	74	9	-E07	come dire	154	2	-E11

Tabella 4. Forme testuali specifiche e rare per città

5.1. *Milano*

Dall'analisi delle parole chiave delle interviste condotte a Milano (Tabella 4) emergono tutti i temi in studio, descritti con linguaggio competente: ad esempio, "interculturale" e "multiculturalità". L'atteggiamento è essenzialmente positivo, emerge una grande attenzione, anche delle istituzioni, per i temi sociali, che vengono trattati in modo concreto, per mezzo di progetti orientati alla diffusione delle informazione, anche mediante l'uso dei nuovi mezzi di comunicazione, all'inserimento lavorativo e culturale degli immigrati. La scuola svolge un ruolo fondamentale in questi processi. Parole chiave sono "insegnanti", "educazione alla sessualità", "progetto", "soldi", "politica", "esperti", "mercato del lavoro", "comportamenti a rischio", "alfabetizzazione", "alunni stranieri", ed in maniera pressoché originale "internet" e

"computer". Nonostante sia tipica la forma "solidarietà", gli immigrati vengono definiti "loro", sono sempre considerati una "minoranza". Assente è l'approccio agli argomenti di carattere morale: "associazioni di volontariato", "accoglienza", "coppie omosessuali", "religione", sono tutte forme rare. Inoltre, l'immigrazione non è di primo arrivo, come testimonia l'assenza delle forme "rifugiati", "clandestini", "accoglienza". Milano, a detta degli esperti, è dunque una città che affronta le proprie tematiche sociali in maniera molto pratica, risolutiva, con grande interesse e partecipazione delle istituzioni. L'aspetto etico viene tenuto in secondo piano, probabilmente lasciato al privato. Se entriamo nel merito dell'analisi delle specificità nell'ambito delle singole sezioni tematiche possiamo approfondire ulteriormente le peculiarità dell'approccio degli esperti milanesi. Sia dall'assenza di terminologia specialistica e che dagli argomenti oggetto delle risposte emerge una complessiva genericità nel trattamento di argomenti che invece presentano precise caratteristiche e modalità di analisi. Questo si ravvisa in particolare nella parte relativa ai temi inerenti la popolazione e le sue dinamiche, dove non appaiono parole o segmenti riferiti specificamente a tali problematiche, se non nell'accezione generale di "tendenze demografiche", bensì argomenti con connessioni o collegamenti indiretti o tangenziali alle problematiche demografiche (rapporto di genere, educazione alla sessualità, educazione alimentare, educazione interculturale), o con deboli legami (fumo, salute, droghe, computer). D'altro canto si evidenzia una forte attenzione alle ricadute didattiche di tali argomenti (formazione degli insegnanti, aggiornamento, programmi). Per quanto riguarda le trasformazioni in ambito familiare sembra prevalere tra gli esperti milanesi una maggiore consapevolezza dei meccanismi alla base dei mutamenti in corso ("fanno meno figli", "ritardo") ma semplificata e limitata alle problematiche di maggior impatto, senza entrare nel merito della complessità e diversificazione delle nuove forme familiari che indubbiamente caratterizzano la società milanese. Si nota inoltre una accentuazione sugli aspetti conflittuali ed etici del tema famiglia (responsabilità, dei grossi problemi), più che su quelli sociali.

L'impatto dell'immigrazione è percepito e vissuto attraverso i suoi aspetti più appariscenti: quello della comunità "cinese" che a Milano è la più grande d'Italia, quello della "criminalità" e quello della integrazione ("integrati"); specifico del mondo della scuola (insieme a "mediazione culturale") è il tema della "loro lingua", pur nell'accezione problematica dell'alterità, e delle "difficoltà di inserimento". Nella consapevolezza legata alle caratteristiche del mercato del lavoro del nord-est ("risorse", "lavori che gli italiani non fanno") compare anche il tema della funzione produttiva dell'immigrazione, pur sottostante alla necessità di "limitare". Anche sul tema dei ruoli di genere si riscontra una complessiva genericità di approccio, in assenza di parole chiave specifiche, è forte anche in questo caso il richiamo al legame tra la scuola e la famiglia (educazione dei figli), nel solco della "tradizione" e dell'"esperienza". L'immagine delle prospettive dei giovani nella società e nel mondo del lavoro proiettata dal linguaggio degli esperti è fondamentalmente legata al presente: tra il divertimento e l'impegno sociale e politico, a sostegno della pace, in cui la dimensione lavorativa ha un ruolo marginale, rispetto a tematiche forti che vanno dalla religione ai problemi delle donne, dove è importante la realizzazione di se stessi.

5.2. *Venezia*

Dal linguaggio utilizzato nei testi delle interviste condotte a Venezia emerge una minore sicurezza tecnica rispetto agli argomenti trattati: gli intervistati tendono a rendere più soggettiva e meno istituzionale la riflessione sulle tematiche proposte, ad accentuare cioè i propri punti di vista: "secondo me", "mi pare", "non so", "vita quotidiana". Come Milano,

Venezia è una città molto attiva nei confronti delle emergenze sociali, ma vi si nota un approccio diverso nelle modalità e negli interventi per affrontarle. Infatti "servizi", Centro Donna", "mediatori culturali", "associazioni di volontariato", oltre a "politiche sociali" danno conto di una attività dell'Ente locale dedicata alle tematiche sociali emergenti. Marginale sembra il ruolo della scuola anche se emerge una specifica attenzione al mondo dell'infanzia e dell'adolescenza. Anche in questo caso si nota un certo pregiudizio nei confronti di "queste persone" e di "il diverso". Gli immigrati sono essenzialmente "profughi", "rifugiati" e "clandestini", per la maggior parte provenienti dai paesi della ex-Jugoslavia. Emerge un quadro di Venezia come municipalità molto attiva nel settore sociale, dove gli aspetti 'tecnici' della gestione del sociale appaiono meno rilevanti di quanto è invece emerso a Milano. Nel merito dei singoli argomenti notiamo che il linguaggio degli esperti di Venezia si caratterizza per una maggiore accuratezza nella scelta dei temi citati che delle parole selezionate, provenienti da un vocabolario preciso e competente. Nella sezione dedicata alla demografia troviamo infatti termini come "invecchiamento della popolazione", "anziani", "natalità", "politiche sociali", "dinamiche demografiche", che appartengono per significato e terminologia al linguaggio di chi possiede una buona competenza della materia. Come abbiamo già notato queste competenze si intrecciano con l'attività istituzionale (servizi, operatori, politiche sociali, assessorato, centro donna, programmazione) confermando un atteggiamento di forte vocazione di servizio connotata dalla soggettività dei parlanti (non so, non mi pare). Anche nel campo delle trasformazioni della famiglia appaiono termini che rimandano ad una visione ampia della problematica che include gli aspetti più recenti e complessi del nuovo panorama familiare e affettivo (coppie omosessuali, famiglie monogenitoriali, al di fuori del matrimonio, un solo figlio), pur in una percezione preoccupata e problematica (problemi, stress). Il tema dell'immigrazione è impregnato della recente e traumatica esperienza della città a confine con il conflitto nella ex Jugoslavia (Jugoslavia, serbo, rifugiati, Sarajevo) e delle iniziative di risposta all'emergenza (campi profughi, accoglienza, mediatori linguistici e culturali), dove appare sempre il contributo dell'amministrazione e dunque il forte spirito di servizio degli intervistati. Ovviamente emergono anche i problemi connessi all'immigrazione (clandestini, prostituzione) ma sempre in una prospettiva attiva di attenzione civica (dignità, tessuto sociale, protezione sociale). Infine il tema dei ruoli di genere appare trattato in un approccio coerente con i precedenti, ossia metabolizzato attraverso competenza ed esperienza, così suggerisce l'uso di termini che contemplano una percezione del tema piuttosto approfondita ed elaborata (flessibilità, identità, modelli, condivisione), ancora una volta la tematica appare collegata ad interventi ed iniziative dell'istituzione municipale (asili nido, pari opportunità, programmazione) e delle istituzioni scolastiche (educazione sessuale, anche nella scuola, la scuola dovrebbe, carenze, qualche insegnante).

5.3. *Salerno*

I soggetti intervistati a Salerno affrontano la trattazione degli argomenti proposti rifacendosi spesso a riflessioni personali, in modo molto discorsivo e poco tecnico: "penso che", "ritengo che", "credo che" sono forme specifiche. Parametri di valutazione sono la "mentalità" "tradizionale", la "religione" e il "senso di responsabilità". La scuola svolge un ruolo importante per l'"informazione" e per "insegnare", "nei programmi", scolastici, appunto. La scarsa progettualità emerge anche dall'uso raro delle forme "progetto", "soldi", "servizi", "sostegno". Salerno, dunque, è una città che tenta un'apertura culturale, ma le competenze sono ancora da formare, le istituzioni devono ancora attivarsi e dove il mondo della scuola ed il sistema educativo rappresentano poli importanti di aggregazione e attenzione. Entrando nel dettaglio delle specifiche sezioni tematiche possiamo confermare questa prima valutazione. Nei

colloqui dedicati ai temi demografici in generale troviamo citati due fenomeni caratteristici (invecchiamento, calo demografico) che denotano una corretta seppur basica — percezione delle principali tendenze in atto, strettamente collegata al "mondo della scuola", ai suoi" istituti"ai suoi "docenti", pur nella consapevolezza della difficoltà ("non credo che i giovani") di "affrontare o "proporre" questo "argomento", o di vederlo integrato "programmi" "ministeriali". Anche sul tema delle nuove famiglie ci troviamo di fronte ad un atteggiamento che contempla diversi aspetti del problema, dalle "coppie omosessuali" al "nucleo familiare" "tradizionale" all'"avere un figlio" "fuori dal matrimonio", secondo un approccio basato sul "senso di responsabilità", sull'esigenza di "serenità" e di un "punto di riferimento", sulla consapevolezza della "sofferenza" di coloro che "sono diversi". Una modalità dunque più fondata sul buon senso e l'esperienza che sulla scorta di specifiche competenze, o di applicazione di politiche, interventi o servizi. Ancora sull'immigrazione si conferma un approccio che è una miscela di consapevolezza, tradizione e istanze istituzionali (soprattutto la scuola "la scuola ha un ruolo", "provveditorati"). Si avverte anche una originale attenzione al tema della religione (l'ora di religione, storia delle religioni, insegnamento della religione), mentre, come nelle altre città, il fenomeno migratorio provoca sentimenti contrastanti che vanno dalla necessità di "accogliere", di "una società multiculturale", al bisogno di "una apertura mentale", per arrivare anche agli aspetti più conflittuali ("criminali", "delinquenza", "clandestino"). Un altro elemento importante è quello della consapevolezza del "noi abbiamo bisogno" di "manodopera", del riconoscimento dell'immigrazione come "risorsa", ed in particolare nel settore dei servizi alla persona ("badanti") che va connesso al peculiare innesto dei lavoratori stranieri nel contesto produttivo della Campania. Nella trattazione del tema dei ruoli di genere prevale ancora una volta una visone privata piuttosto che una logica pubblica di servizio o sociale, da giocare all'interno della dimensione familiare e coniugale ("casa non", "collaborare, "collaborazione", "mentalità"), da affrontare attraverso strumenti come "l'educazione dei figli" o il "parlarne", mentre si conferma il ruolo centrale dell'istituzione scolastica "la scuola può".

6. Analisi delle corrispondenze lessicali

6.1. *Le categorie di interviste*

Per analizzare l'eventuale esistenza di differenze di atteggiamenti e di approccio in base alla città di appartenenza, compresa la classica differenza di mentalità e di organizzazione (anche politica-istituzionale) tra nord e sud Italia, nei confronti dei temi analizzati, sulla base dei contenuti espressi nel corso delle interviste, le sequenze di testo sono state riclassificate in base alla città di appartenenza del soggetto intervistato e alla tematica cui si riferiscono, ovvero alla sezione del questionario[i]. L'analisi è stata effettuata comparando i sub-testi ottenuti da questo nuovo raggruppamento. Il primo fattore si riferisce all'*area tematica*, alla sezione dell'intervista. Interessante è notare la vicinanza tra i testi delle interviste che trattano il tema dei ruoli di genere e quelli sulla famiglia: in realtà i due argomenti sono stati percepiti come molto simili per contenuti. Sul secondo asse le interviste sono distribuite in base alla città di appartenenza dell'intervistato: Venezia, Milano, Salerno; il fattore si riferisce pertanto all'*elemento geografico*. Mentre le interviste condotte a Milano si collocano in una posizione intermedia, verso il baricentro del grafico, quelle di Salerno e Venezia sono del tutto opposte per contenuti. Risalta inoltre l'originalità delle interviste dei veneziani che si distinguono nettamente da quelle delle altre due città che presentano alcuni elementi di linguaggio comune.

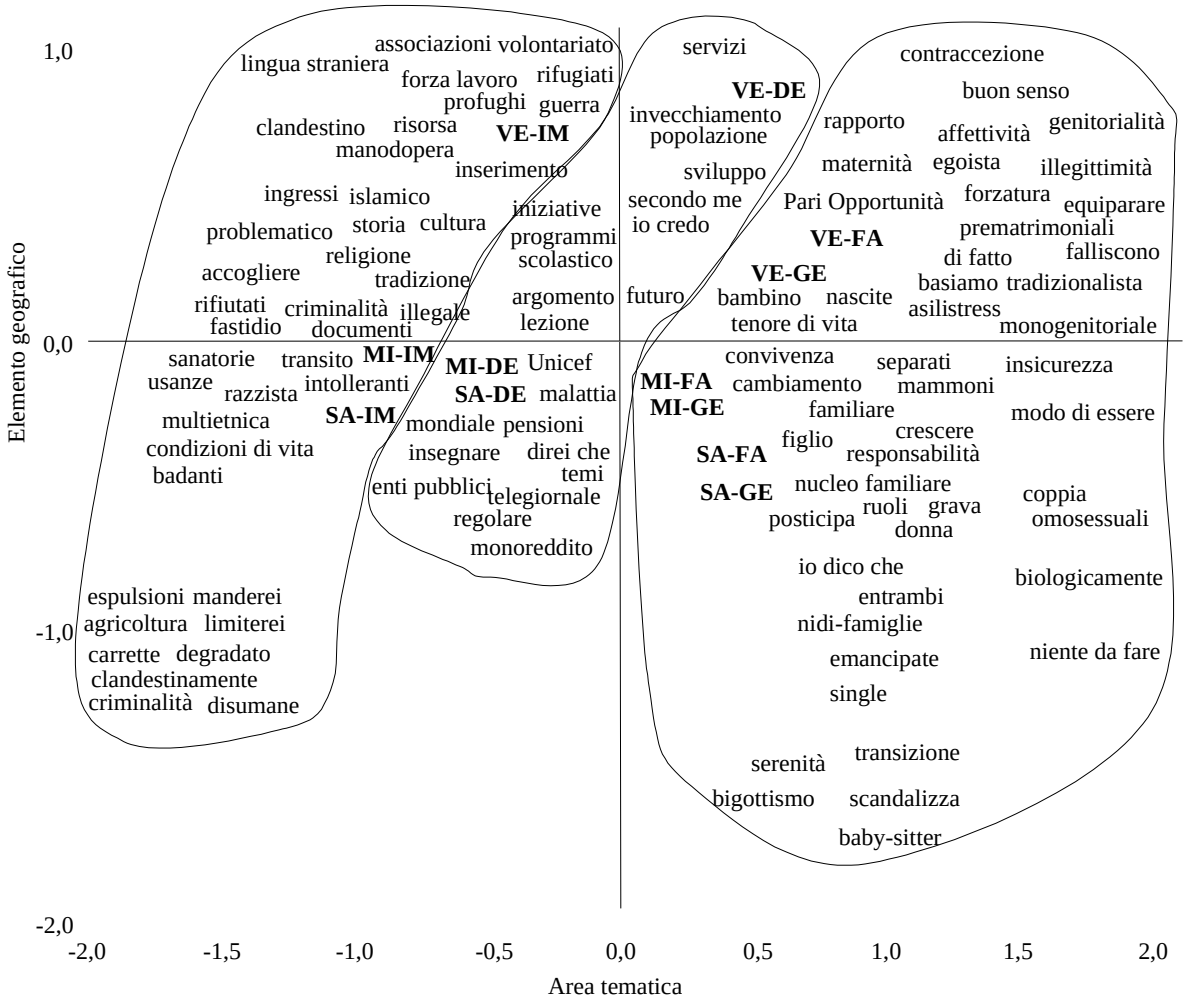


Grafico. Le aree tematiche. Piano fattoriale 1-2

6.2. *Le aree tematiche*

L'analisi dei risultati è stata effettuata comparando le tre città rispetto alle sezioni del questionario, i "grandi temi" oggetto dello studio.

6.2.1. *Immigrazione*

Il fenomeno dell'immigrazione a Venezia è dominato dalla presenza di immigrati provenienti dalla ex-Jugoslavia, a causa della guerra e della prossimità geografica. Presenti sono riferimenti a "rifugiati", "profughi", "clandestini"; di coloro si occupano "associazioni" di "volontariato". Gli stranieri residenti sono inseriti nel tessuto economico e sociale della città: sono una "risorsa", "forza lavoro", e c'è interesse per la loro "cultura" e la loro "storia". Molto interessante appare la presenza della dimensione dell'integrazione culturale, legata alla "storia", alla "cultura" ed alla "tradizione" minacciata dalla "problematicità"connessa alla "accoglienza" di persone di "religione" "islamica". "Problematico", anche, è considerato l'ingresso di 'irregolari', che vengono associati alla "criminalità", sono "rifiutati". Altrettanto interessante appare la reazione istituzionale che propone "iniziative" e "programmi scolastici" che introducano l'immigrazione come "argomento di lezione". In sintesi possiamo dire che per gli esperti veneziani il tema si declina secondo tre assi di lettura: quello dell'esperienza legata al recente conflitto balcanico, quello della gestione del fenomeno attraverso una radicata cultura di servizio e di intervento, quello più ampio del suo impatto culturale. Per gli intervistati di Milano l'immigrazione non è di 'arrivo', bensì di "transito"; affiora anche

l'approccio pragmatico sintetizzato dallo strumento delle "sanatorie", mentre emerge una certa preoccupazione dovuta all'intolleranza, causata da un atteggiamento razzista. Più articolata, e molto vicina alle posizioni degli intervistati milanesi, è la posizione degli esperti di Salerno, che dichiarano una decisa avversione all'immigrazione clandestina, quella che, nel meridione d'Italia si realizza per mezzo delle cosiddette "carrette" del mare. Per queste persone gli intervistati desidererebbero "espulsioni", condannando le condizioni "disumane" e di degrado in cui si trovano a vivere gli immigrati clandestini. Auspicano anche una limitazione dei nuovi ingressi. Assieme a questa dimensione troviamo un approccio più positivo suggerito dalla necessità di una società "multietnica", dalla preoccupazione per le "condizioni di vita" degli immigrati, dal ruolo delle badanti nella società italiana.

6.2.2. *Tendenze demografiche*

Il tema popolazione si sviluppa, graficamente, intorno al centro del piano fattoriale ed in prossimità degli assi: non è molto caratterizzante, ed infatti è l'argomento sul quale gli intervistati si sono dichiarati poco competenti, per cui dai colloqui non sono emersi approfondimenti del tema. Come abbiamo già notato, gli intervistati di Venezia trattano il tema secondo due direttrici, una più locale connessa all'invecchiamento della popolazione, ponendo l'attenzione sui "servizi", l'altra in una dimensione più globale laddove emerge il tema del "futuro" e dello "sviluppo", che vengono connessi alle tematiche demografiche. A Milano, il tema è trattato con un approccio meno teorico e più operativo: numerose sono le "iniziative" o "programmi" a livello "scolastico" attuate o che si vorrebbero realizzare per divulgare le informazioni sull'argomento. Informazioni che a Salerno arrivano tramite "telegiornale", e che interessano per il risvolto pratico, quello delle problematiche legate alla vita delle persone che vivono di "pensioni" o affette da "malattia", dalle famiglie "monoreddito", appare peraltro anche la dimensione istituzionale legata al ruolo degli "enti pubblici" o delle organizzazioni che operano a livello "mondiale" come l'"Unicef".

6.2.3. *Famiglia, nuove famiglie e ruoli di genere*

La sovrapposizione (o la confusione) delle problematiche dei ruoli di genere con quelle legate alla trasformazione delle famiglia è di per sé una prima indicazione della mancanza di chiarezza su due sfere tematiche in realtà ben distinte. Si tratta di un aspetto che caratterizza tutti gli esperti intervistati, a prescindere dalla città di appartenenza. Le tematiche di genere sono tuttora percepite all'interno di un'ottica familiare e non vengono proiettate in un orizzonte più ampio che è quello sociale dei ruoli culturalmente accreditati ai due generi, si tende dunque a recepire solo gli aspetti più tradizionali, senza entrare nel merito delle specifiche modalità del dibattito culturale e politico più recente. Gli argomenti sono affrontati, dagli intervistati di Venezia, in un'ottica che contempla sia istanze di gestione 'pubblica' ("contraccezione", "asili" "Pari opportunità", "equiparare" famiglie "di fatto"), che uno sforzo interpretativo (la "genitorialità", l'"affettività", "maternità", l'attenzione per i bambini, al "buon senso" nel decidere il modello "educativo", "stress") . Si esprime "biasimo" verso le tipologie familiari alternativa a quella "tradizionalista", che generano "stress" e "falliscono". Emerge anche una capacità di lettura delle linee di intervento sui modelli di famiglia, quello tradizionalista, le unioni prematrimoniali, o quelle monogenitoriali. A Milano sono sintetiche le parole chiave che emergono dall'analisi — attestandosi le testimonianze anche questa volta in posizione intermedia-, ma in grado di indicare l'acquisizione dei principali elementi problematici — pur in un'ottica prettamente familiare, dove la lettura di genere non compare. Si prende innanzitutto atto del "cambiamento" in corso, dell'aumento delle "separazioni" e dell'"insicurezza", della permanenza dei giovani nelle famiglie di origine ("mammoni"). Anche a Salerno l'attenzione degli esperti è polarizzata su tematiche prettamente familiari, in un'ottica prevalentemente tradizionale: la "responsabilità" di "crescere" i figli "grava" sulla "donna", i "ruoli" non sono cambiati. Appare anche l'inclusione di temi controversi o complessi come quello delle coppie "omosessuali", per le quali tuttavia non c'è "niente da fare",

"biologicamente". Con le parole degli esperti, la "transizione" verso modelli si attraversa nella società salernitana tra "serenità"e "bigottismo".

7. Conclusioni

In conclusione, sul piano del superamento degli stereotipi, dei pregiudizi e delle resistenze culturali c'è ancora molta strada da fare. Un segnale è dato dal fatto che tutte le tematiche proposte sono state percepite dagli intervistati innanzitutto come "problema" di cui occuparsi e non semplicemente come realtà sociali emergenti. Le dinamiche demografiche toccano solo di sfuggita l'interesse dei soggetti intervistati. I testimoni, infatti, si sono limitati ad esprimere opinioni impregnate di "senso comune", senza che nessuno si allontanasse dal campo del "sentito dire". Le opinioni sull'immigrazione, invece, li hanno coinvolti. E' evidente una tendenza all'apertura ed all'accettazione ancora acerba; molti degli stereotipi che circolano sugli stranieri sono stati in parte confermati dall'indagine, tuttavia, è evidente il desiderio di cambiare atteggiamento che si evince dal favore mostrato riguardo all'introduzione dello studio dell'immigrazione a scuola. La sezione dell'inchiesta relativa agli atteggiamenti sulle nuove famiglie ha visto emergere tutti gli elementi della radicale trasformazione che ha investito la famiglia italiana, in direzione della riduzione dei matrimoni, della contrazione della natalità, della preferenza verso la convivenza e le unioni libere. Anche nel caso delle nuove famiglie, emerge una indiscussa preferenza per la configurazione familiare classica che resta un modello intramontabile, l'unico in grado di assicurare benessere a tutti i suoi membri, in particolare ai figli. Questa osservazione è valida anche per quanto attiene alla ridefinizione delle identità dei due generi: sono ancora validi i ruoli tradizionalmente e socialmente destinati ai due sessi, che restano oggi sostanzialmente invariati nella pratica, al di là di qualunque rivendicazione femminile. La persistenza di un modello culturale fortemente ancorato alla tradizione si è mostrata in tutta la sua evidenza anche quando è stato opposto un "no" deciso alla legittimazione delle coppie omosessuali. Sembrerebbe che gli intervistati — e invero l'opinione pubblica in generale — sono ancora impreparati a tenere il passo del cambiamento. Del resto, è noto che la metabolizzazione culturale avviene ad un ritmo di gran lunga più lento di quello che si verifica nella società.

Bibliografia

Baccaini B. e Gani L. (2002). *La population en questions – Une enquête sur les connaissances et représentation sociodémographiques des élèves de terminale*. INED.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Bolasco S. e Morrone A. (1998). La construction d'un lexique fondamental de polyformes selon leur usage. In *Actes des JADT 1998* : 155-166.

Bolasco S., Morrone A. e Baiocchi F. (1999). A paradigmatic path for statistical content analysis using an integrated package of textual data treatment. In Vichi M. e Opitz O. (Eds), *Classification and data analysis. Theory and application*. Springer-Verlag.

Cavalli A. e De Lillo A. (1993), a cura di, *Giovani Anni 90*. Il Mulino.

Guiraud P. (1960). *Problèmes et méthodes de la statistique linguistique*. PUF.

Misiti *et al.* (2003). *Under 18- Numeri, idee, ideali degli studenti di Venezia. Le tendenze demografiche, la famiglia, i ruoli di genere, gli immigrati*. Venezia – Cittadinanza delle donne e Cultura delle Differenze.

Morrone A. (1993). Alcuni criteri di valutazione della significatività dei segmenti ripetuti. In Anastex, *JADT 1993* : 445-453.

Oecd (1997). *Promoting Public Understanding of Science and Technology*. OECD/GD(97) 52, OECD.

Cartographie automatique du contenu d'un corpus de documents textuels

A. Mokrane, R. Arezki, G. Dray, P. Poncelet

Groupe connaissance – LGI2P – EMA
Parc scientifique Georges Besse, Site EERIE
30035 Nîmes Cedex 1 – France
{abdenour.mokrane, rachid.arezki, gerard.dray, pascal.poncelet}@ema.fr

Abstract

Considering the quantity of documents available nowadays, the user (organization, company, individual, etc.) is overloaded by information. This one is unable to analysis or apprehend this information in their globality. With the Web, the not structured textual documents became prevalent. Useful information being hidden in the text, it becomes essential to propose new systems allowing the analysis, the organization and the representation of the various textual contents. In this paper we propose a new system, called IC (Information Cartography), for the automatic cartography of the contents of textual documents corpus. IC is based on an original approach for the selection of the representative terms of the content. It helps a community of users working on a thematic in its documentary consultations, by proposing an information chart on the total contents of the corpus or/and for each document.

Keywords : Text mining, Information cartography, Contextual co-occurrence, Representative term.

Résumé

La quantité d'informations et de documents disponibles de nos jours, entraîne une « surinformation » de l'utilisateur final (entreprise, organisme, individu, etc.) qui n'est donc plus capable d'analyser ou d'appréhender ces informations dans leur globalité. Avec le Web, les documents textuels non structurés sont devenus prédominants. L'information utile étant enfouie dans le texte, il devient indispensable de proposer de nouveaux systèmes permettant l'analyse, l'organisation et la représentation des différents contenus textuels. Dans cet article nous proposons un nouveau système, appelé IC (Information Cartography), pour la cartographie automatique du contenu d'un corpus de documents textuels. Le système IC est basé sur une approche originale pour le choix des termes représentatifs du contenu d'un corpus documentaire d'une thématique. Il permet d'aider une communauté d'utilisateurs travaillant sur une thématique donnée dans ses consultations documentaires en lui proposant une carte d'information sur le contenu global du corpus et/ou de chaque document. Nous illustrons IC sur un corpus d'articles de presse.

Mots-clés : fouille de textes, cartographie d'information, co-occurrence contextuelle, terme représentatif.

1. Introduction

Chaque jour, le nombre de documents disponibles croît d'une manière exponentielle, en particulier en raison de l'essor des communications électroniques. Nous sommes entrés dans l'ère de l'information où l'utilisateur est submergé par la quantité d'information disponible. Avec le Web, les documents textuels non structurés sont devenus prédominants. Les informations utiles étant enfouies dans les textes. Les utilisateurs ne sont donc plus capables d'analyser ou d'appréhender ces informations dans leur globalité. Pour cela, il est nécessaire de mettre en

œuvre des systèmes permettant d'analyser les contenus des documents, de les organiser et de les représenter automatiquement.

Actuellement, de nombreux travaux de recherche, notamment issus du Web Mining et du Text Mining, s'intéressent à la fouille de corpus de documents textuels volumineux (Poibeau, 2003 ; Besançon, 2001 ; Chen *et al.*, 2001 ; He *et al.*, 2001) Ces travaux ont donné naissance à des systèmes de catégorisation, voire de cartographie de documents tel que Kartoo (Chung *et al.*, 2002) ou Mapstan (Spinat 2002). Ces outils retrouvent des liens entre les différents documents ou sites Web et représentent ces liens sous forme de carte de navigation. Cependant les informations du contenu du corpus documentaire ou de chacun des documents sont peu représentées.

L'objectif de ce papier est de présenter un système, IC (Information Cartography), de cartographie automatique du contenu d'un corpus de documents textuels, basé sur une approche originale pour la sélection des termes représentatifs du contenu. L'article est organisé de la manière suivante. La section 2 résume les travaux antérieurs liés à notre problématique. La section 3 décrit l'architecture fonctionnelle du système IC. Dans la section 4 nous détaillons les différentes étapes de cartographie de contenus textuels. La section 5 illustre la mise en œuvre de notre approche sur un corpus d'articles de presse. Enfin, la section 6 conclue ce papier et présente les perspectives de recherche associées.

2. Travaux antérieurs

Les outils et les méthodes de fouille de textes permettent l'acquisition, le classement, l'analyse, l'interprétation, l'exploitation et la visualisation systématiques d'informations contenues dans des documents textuels (Poibeau, 2003). Actuellement, de nombreux travaux de recherche, notamment issus du Web Mining (Kosala et Blockeel, 2000) et du Text Mining, s'intéressent à la fouille de corpus documentaires volumineux (Andrieu, 2000 ; Besançon, 2001 ; Chen *et al.*, 2001 ; Han et Kamber, 2000 ; He *et al.*, 2001 ; Turenne, 2000). L'objectif de ces travaux est généralement d'analyser le contenu des documents pour en extraire des termes significatifs ainsi que les liaisons qui peuvent exister entre ces différents termes. Dans ce cadre, les modèles de similarités textuelles et la notion de co-occurrences sont les plus utilisées pour l'analyse du contenu (Poibeau, 2003). Dans un contexte proche, celui de la recherche documentaire, la recherche de co-occurrences a également été largement étudiée ces dernières années, elle consiste à rechercher les associations de termes les plus fréquentes dans les documents afin de retrouver rapidement les documents pertinents qui peuvent répondre aux requêtes de l'utilisateur. Dans Pereira *et al.* (1993) cette co-occurrence est utilisée pour la classification des termes selon la distribution de leurs contextes syntaxiques. Tanaka *et al.* (1996) utilisent la matrice de co-occurrences pour la désambiguïsation des termes. Besançon *et al.* (2002) ont proposé un modèle de filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents et la recherche documentaire. Tous ces travaux de fouille de textes, ne prennent pas en considération les co-occurrences dans le choix des termes de l'espace vectoriel ou les termes représentatifs du contenu d'un corpus documentaire. Ce qui implique une pénalisation d'une partie importante des relations de co-occurrences. Cette pénalisation est due au choix de termes basé principalement sur leurs fréquences d'occurrences (Salton *et al.*, 1975 ; Salton et Buckley, 1988). L'application de ces approches pour la cartographie du contenu des documents est donc limitée dans la mesure où elle ne permet pas d'extraire des informations pertinentes sur le contenu des documents. En effet, ces méthodes ne contiennent pas de stratégie, prenant en considération les associations

des termes et leurs occurrences, pour sélectionner les termes représentatifs d'un corpus documentaire d'une thématique donnée.

3. Architecture fonctionnelle du système IC

Le système IC (Information Cartography) se charge de la cartographie du contenu d'un corpus de documents textuels. Il est composé de trois sous-systèmes dédiés aux trois étapes de notre méthodologie de cartographie du contenu : analyse linguistique et prétraitements, analyse statistique des données textuelles, modélisation et réalisation des cartes d'information. La figure 1 illustre l'architecture fonctionnelle du système IC.

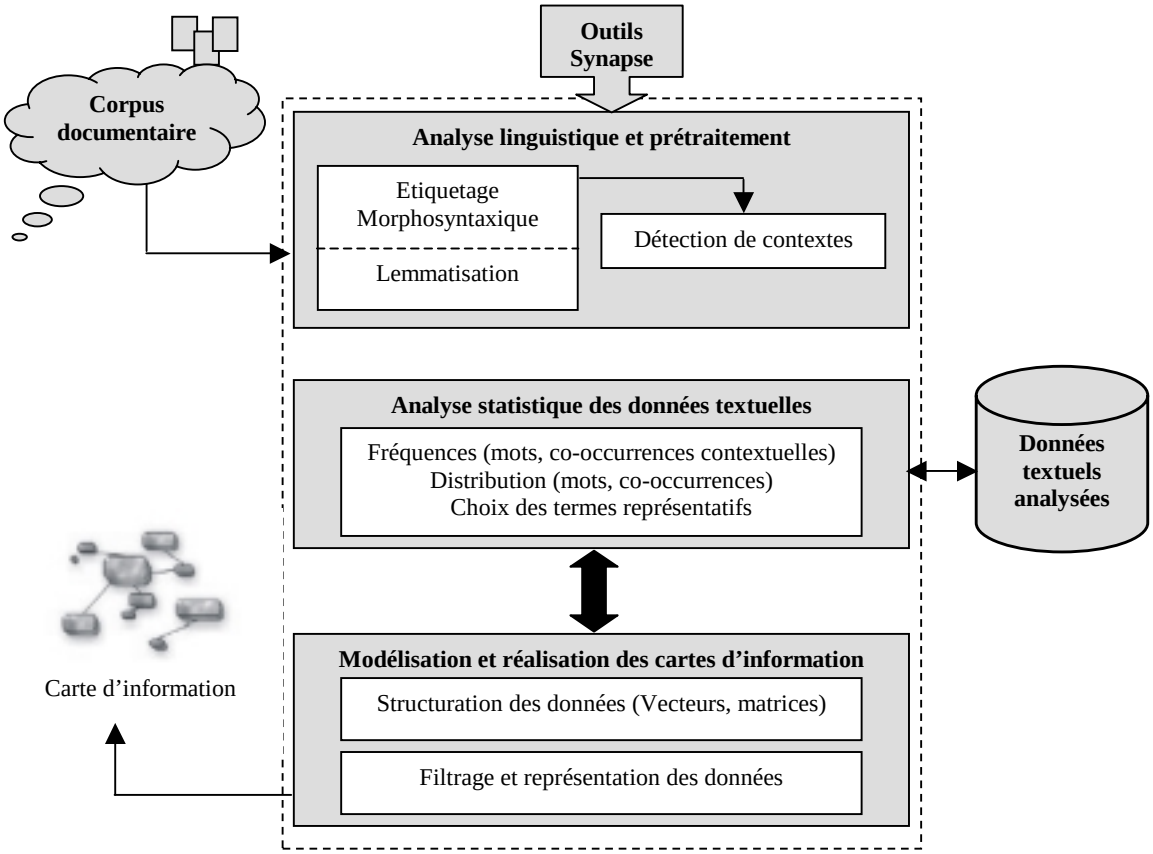


Figure 1. Architecture fonctionnelle du système IC

4. Fonctionnement du système IC

Etant donné que nous ne nous intéressons pas ici au traitement automatique du langage naturel (TALN), nous utilisons l'analyseur de la société Synapse (http://www.synapse-fr.com) qui intègre un étiqueteur morphosyntaxique et un lemmatiseur fonctionnant pour les documents textuels en Français. Via ces deux outils, le corpus de documents textuels est transformé en documents étiquetés et lemmatisés. Le prétraitement concerne le nettoyage du corpus des mots vides (articles, pronoms, prépositions, etc.) et la détection des différents contextes. À l'aide des étiquettes, nous conservons principalement les noms, les verbes et les adjectifs.

4.1. *Analyse statistique des données textuelles*

Comme nous l'avons spécifié dans la section travaux antérieurs, la notion de co-occurrence, prend des dimensions différentes suivant la tâche à accomplir. Pour la recherche d'informa-

tion, on définit généralement des co-occurrences entre des mots et des documents contenant ces mots. Pour la détection des contextes de termes, la co-occurrence est entendue entre le terme et une fenêtre de mots. Enfin une autre approche consiste en l'acquisition de co-occurrences entre les différentes unités d'un corpus de référence. Ce type d'approche a été adopté par exemple dans le cadre du modèle DSIR (Distributional Semantic for Information Retrieval) (Besançon 2002). Le modèle DSIR intègre la notion de co-occurrences entre termes pour une nouvelle représentation vectorielle de documents. Cependant, le modèle proposé, ne prenant pas en considération les co-occurrences dans le choix des termes de l'espace vectoriel, implique une pénalisation d'une partie importante des relations de co-occurrences. Cette pénalisation est due principalement au choix des mots clés basé sur leurs fréquences d'occurrences. Pour répondre à cette problématique, nous proposons dans la suite de cette section, une nouvelle stratégie pour établir le choix des termes représentatifs du contenu global ou de chacun des documents. Cependant, dans un premier temps, nous adaptons les définitions de co-occurrence à notre contexte.

4.1.1 *Définitions*

De manière générale, dans les différentes approches existantes, un contexte de co-occurrence peut être une phrase, un paragraphe ou même l'ensemble du document. Etant donné que nous considérons que les éléments pertinents sont généralement proches dans un document, nous considérons dans notre modèle qu'un contexte correspond à une phrase et ainsi la détection des contextes va correspondre à l'annotation des différentes phrases du corpus documentaire. Dans ce cadre, nous pouvons définir la notion de co-occurrence contextuelle de la manière suivante :

Co-occurrence contextuelle (*CO*) : Deux termes *A* et *B* appartenant, en même temps au même contexte, forment une co-occurrence appelée *CO* et notée *{CO : A—B}*.

Pour chaque terme, il est possible de connaître d'une part sa fréquence d'apparition dans le corpus mais également le nombre de documents qui contiennent ce terme.

Fréquences d'un terme (*FTC* et *FTD*) : La fréquence *FTC* d'un terme *T* dans un corpus de documents textuels correspond au nombre d'occurrences du terme *T* dans le corpus. La fréquence *FTD* d'un terme *T* dans un corpus de documents textuels correspond au nombre de documents contenants *T*.

La définition précédente est bien entendue spécifique à la notion de terme et les définitions suivantes sont adaptées à la notion de co-occurrence.

Fréquences d'une co-occurrence (*FCC* et *FCD*) : La fréquence *FCC* d'une co-occurrence *CO* dans un corpus de documents textuels correspond au nombre d'occurrences de *CO* dans le corpus. La fréquence *FCD* d'une co-occurrence *CO* dans un document *D* correspond au nombre d'occurrences de *CO* dans *D*.

À partir des différentes fréquences obtenues, il est possible de créer les matrices suivantes :

Matrice de co-occurrences brute (*MATCO*) : Soit *N* le nombre de termes d'un corpus documentaire. La matrice de co-occurrence brute notée *MATCO* correspond à une matrice de *N* ligne et *N* colonnes. La ligne *i* de la matrice correspond à un terme *Ti* du corpus et la colonne *j* de la matrice correspond à un terme *Tj* du corpus (*i= 1..N, j= 1..N).*

$$Si\ (i \neq j) \quad MATCO\ (i,j)=FCC \text{ de } \{CO : Ti—Tj\} \quad \text{Sinon} \quad MATCO\ (i,j)=FTC\ de\ Ti$$

La première analyse statistique consiste à calculer les *FTC, FTD, FCC et MATCO* (définies ci dessus) dans l'objectif de choisir les termes représentatifs. À partir de la matrice *MATCO* et des termes représentatifs du corpus documentaire, nous construisons la matrice de co-occurrences réduite définie de la manière suivante :

Matrice de co-occurrences réduite (*RMATCO*) : Soit *E* l'ensemble des termes d'un corpus documentaire et considérant l'ensemble $R \subset E$, contenant *M* termes.

La matrice de co-occurrences réduite notée *RMATCO* correspond à une matrice de *M* ligne et *M* colonnes. La ligne *i* de la matrice correspond à un terme *Ti* de l'ensemble *R* et la colonne *j* de la matrice correspond à un terme *Tj* de l'ensemble *R*. (*i= 1..M, j= 1..M).*

Si (i ≠ j) RMATCO (i,j)=FCC de {CO : Ti—Tj} sinon RMATCO (i,j)=FTC de Ti

Nous expliquons dans la section suivante comment nous sélectionnons les termes représentatifs.

4.1.2. *Choix des termes représentatifs*

Le choix des termes représentatifs du contenu du corpus prend en considération les co-occurrences contextuelles les plus fréquentes en plus des fréquences de termes, en tenant compte de la distribution *FTD* de chacun des termes. L'ensemble *E* des termes représentatifs du corpus est constitué par l'union des deux ensembles *E1* et *E2* : $E = E1 \cup E2$.

E1 est l'ensemble des termes *Ti* dont la fréquence d'occurrence dans le corpus *(FTCi)* pondérée par le nombre de documents contenant *Ti* (*FTD)* est supérieure à un seuil α. Dans notre modèle, le seuil α est égal à la moyenne des *FTCi* des termes *Ti* pondérées par les *FTD.* Ce qui se traduit par ceci :

$$\left(FTC_i / FTD \right) > MoyFTC$$

où *MoyFTC* est calculée de la manière suivante : soit *T* l'ensemble des termes du corpus documentaire et *M* le nombre de ces termes. Nous calculons la moyenne pondérée *MoyFTC* des *FTCi* de chacun de ces termes avec prise en considération des *FTD* (nombre de documents contenant chaque terme).

$$MoyFTC = \frac{\sum_{i=1}^{i=M} \left(FTC_i / FTD \right)}{M}$$

E2 est l'ensemble des couples de termes *Ti, Tj* du corpus qui forment une co-occurrence contextuelle *CO* dont la fréquence *(FCC)* est supérieure à seuil β. Dans notre modèle, le seuil β est égal à la moyenne des *FCC* des co-occurrences contextuelles *CO* du corpus documentaire. ce qui se traduit par ceci :

$$FCC \text{ de } \{CO : Ti—Tj\} > MoyFCO$$

Où *MoyFCO* est calculée par : soit *N* le nombre de *CO* du corpus. Nous calculons la moyenne pondérée *MoyFCO* des *FCCi* de chacune des co-occurrences du corpus.

$$MoyFCO = \frac{\sum_{i=1}^{i=N} FCC_i}{N}$$

Après avoir construit l'ensemble *E* des termes représentatifs (union des deux ensembles *E1 et E2*), nous pouvons procéder à la construction de la matrice *RMATCO* (Cf. définitions) sur l'ensemble *E.*

4.2. *Modélisation et réalisation des cartes d'information*

La carte d'information représentant le contenu général du corpus documentaire n'est qu'une représentation sous forme d'un graphe des informations pertinentes de la matrice *RMATCO*. Les sommets du graphe modélisent un ensemble de termes représentatifs et les liens représentent les co-occurrences contextuelles entre ces termes avec leurs degrés d'importance dans le corpus documentaire. Les sommets du graphe sont représentés avec des tailles différentes correspondant aux fréquences des termes dans le corpus. Les différents liens entre les termes sont représentés par des arrêtes de largeurs différentes. Les cartes associées à chaque document du corpus peuvent être construites avec le même modèle appliqué au corpus documentaire.

5. Illustration

Dans le cadre du projet CARICOU (Capitalisation de Recherches d'Informations de Communautés d'Utilisateurs) mené par le LGI2P, nous avons développé l'outil *CO2MAT (COntextual CO-occurrence MATrix)* permettant de construire la matrice *RMATCO qui* stocke les informations reliées à la carte représentant le contenu global du corpus documentaire. Afin de montrer l'intérêt du système IC, nous avons appliqué notre approche via l'outil *CO2MAT* sur un corpus documentaire d'environ 1000 documents de presse. Le corpus documentaire porte sur la thématique « crise irakienne ». Ces documents sont issus de journaux internationaux avant le déclenchement de la deuxième guerre en Irak.

La première étape du choix des termes représentatifs consiste à générer l'ensemble des termes en ne prenant en considération que les fréquences *FTC* des termes dans le corpus général et le nombre de documents *FTD* contenant chaque terme. L'ensemble *SE1* des termes générés par la première étape et non détecté par la seconde ainsi que les termes les plus fréquents de cette étape est le suivant :

SE1 = {Guerre, Irak, Américain, Bush, Faire, Aller, Saddam, Président, Irakien, **France, ONU**}

La deuxième étape consiste à extraire des termes représentatifs en prenant en considération les fréquences de co-occurrences *FCC*. Les co-occurrences contextuelles les plus fréquentes de l'étape 2 ainsi que celles qui ont participé à l'extraction de nouveaux termes non détectés à l'étape 1 sont décrites dans le *Tableau 1* et sont triées par ordre décroissant.

1	CO : Irak – Guerre	6	CO : Bush – Faire	11	CO : Président – Bush
2	CO : Bush – Guerre	7	CO : Faire – Guerre	12	CO : Président – Saddam
3	**CO : Irak – Etats Unis**	8	CO : Aller – Guerre	13	CO : Irak – Président
4	CO : Irak – Américain	9	CO : Bush – Irak	14	CO : Irak – Irakien
5	CO : Bush – Aller	10	CO : Faire – Irak	**15**	**CO : Irak – Inspection**

Tableau 1. Les co-occurrences contextuelles pertinentes

L'ensemble des termes représentatifs correspondant au tableau 1 est le suivant :

SE2 = {**Etats-Unis**, Irak, Irakien, Guerre, Américain, Bush, Aller, Faire, Président, Saddam, **Inspection**}

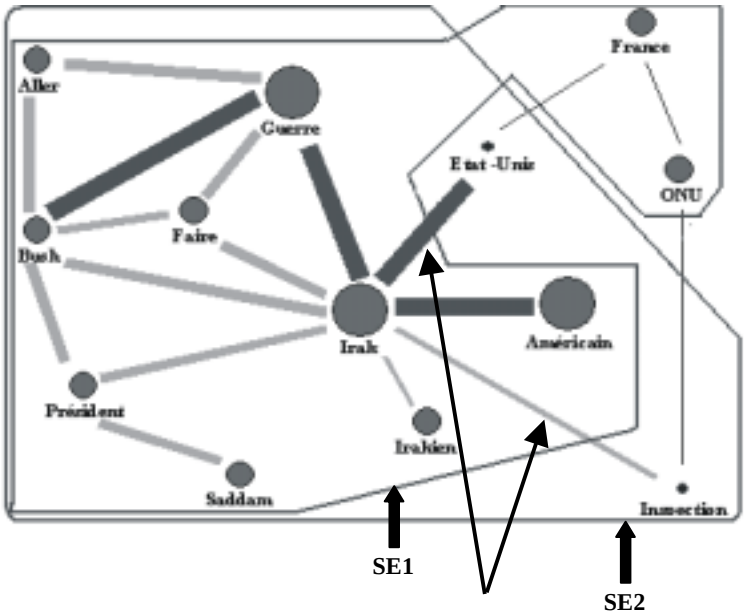
L'ensemble *SE* des termes représentatifs résultant de l'union de *SE1* et *SE2* est donc :

SE= {Guerre, Irak, Américain, Bush, Faire, Aller, Saddam, Président, Irakien, ***France, ONU, Etats-Unis, Inspection***}.

SE1—SE2 = { **France, ONU** }, les termes appartenant à *SE1* et non à *SE2*.

SE2—SE1 = { **Etats-Unis, Inspection** }, les termes appartenant à *SE2* et non à *SE1*.

Nous remarquons que la simple utilisation des fréquences des termes nous prive d'informations importantes sur les Etats-Unis et les Inspections (*SE2—SE1*). La prise en considération des co-occurrences seule mène au même problème, ce qui se traduit par la privation d'informations sur l'ONU et la France (*SE1—SE2)*. Il est clair que les deux étapes de sélection des termes représentatifs sont complémentaires. La figure 2 illustre la carte d'information associée à ces termes représentatifs du contenu, via notre approche.



Co-occurrences contextuelles

Figure 2. Choix et représentation des termes de l'ensemble SE

Dans notre modèle, malgré que le terme Etats-Unis ait une fréquence faible dans le corpus, il a été sélectionné comme terme représentatif. Cela grâce à la co-occurrence contextuelle importante entre Irak et Etats-Unis *{CO : Irak — Etats-Unis}.* De même pour le terme Inspection qui a été sélectionné par rapport à la co-occurrence contextuelle entre Irak et Inspection *{CO : Irak — Inspection}.* Ces nouveaux termes vont permettre de représenter de nouvelles co-occurrences très utiles pour l'interprétation et la représentation du contenu du corpus documentaire (ces co-occurrences ne figurent pas dans le tableau 1). Par exemple les co-occurrences contextuelles entre la France et les Etats-Unis ou bien Inspection et ONU. Ces nouvelles co-occurrences sont faibles dans le corpus mais apportent des informations importantes sur le contenu du corpus documentaire. Notamment en ce qui concerne les inspections en Irak ou bien la relation entre la France et les Etats-Unis, etc. Ces nouvelles co-occurrences apportent de nouvelles informations.

Une première consultation de cette carte, donne une idée général sur le contenu du corpus. On peut déduire rapidement que le corpus documentaire est concentré sur la guerre en Irak. On peut déduire aussi que la presse écrite considère le président Bush comme acteur principal des évènements. Il existe des informations sur la relation Etats-Unis / France dans ce corpus documentaire. On en parle beaucoup moins de l'Organisation des Nations Unies (ONU). Il y a aussi des informations sur les inspections en Irak, etc.

6. Conclusion

Dans cet article, nous avons proposé une nouvelle approche pour la cartographie automatique de contenus textuels. Nous avons concrétisé cette approche par le système IC, basé sur notre approche originale de sélection de termes représentatifs. Cette approche permet d'aider une communauté d'utilisateurs, travaillant sur une thématique donnée, dans ses consultations documentaires. Elle peut également être utilisée, dans le cadre d'un système de recherche documentaire, pour le choix des mots clés et la réduction de l'espace vectoriel. Nos travaux en cours portent sur l'exploitation efficace de cette approche par le développement d'un système de navigation par le contenu, à l'aide des cartes d'information, ainsi que l'application d'algorithmes de clustering flou pour la visualisation de l'ensemble des termes représentatifs de corpus documentaires volumineux et divers.

Références

Andrieu O. (2000). *Créer du trafic sur son site Web.* Eyrolles.

Besançon R. (2002). Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents. In *Actes de TALN*.

Besançon R. (2001). *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles des textes*. PhD thesis, Ecole polytechnique Fédérale de Lausanne.

Chung W., Chen H. et Nunamaker J. (2002). Business intelligence explorer : A knowledge map framework for discovering business intelligence on the Web. In *Proceedings of the 36 Hawaii International Conference on System Sciences (HICSS'03)*, Hawaii.

Chen H., Fan H., Chau M. et Zeng D. (2001). MetaSpider: Meta-searching and categorization on the Web. *Journal of the American Society for Information Science and Technology*, vol. (52) : 1134-1147.

Kosala R. et Blockeel H. (2000). Web Mining research : A survey. *SIGKDD Explorations*, vol. (2/1) : 1-15.

He X., Ding C., Zha H. et Simon H. (2001). Automatic topic identification using Webpage clustering. In *Proceedings of 2001 IEEE International Conference on Data Mining*, Los Alamitos, CA.

Han J. et Kamber. M.(2000) . *Data Mining : Concepts and Techniques.* Morgan Kaufmann Publishers.

Poibeau T. (2003). *Extraction automatique d'information : du text brut au Web sémantique.* Hermès.

Pereira F., Tishby N. et Lee L. (1993). Distributional clustering of English words. In *Proceedings of the 31th Meeting of the Association for Computational Linguistics* : 183-190.

Spinat E. (2002). Pourquoi intégrer des outils de cartographie au sein des systèmes d'information de l'entreprise ? In *Colloque Cartographie de l'information : De la visualisation à la prise de décision dans la veille et le management de la connaissance*, Paris.

Salton G. and Buckley C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, vol. (24/5) : 513-523.

Salton G., Yang C.S. et Yu C.T. (1975). A theory of term importance in automatic text analysis, *Journal of the American Society of Information Science*.

Tanaka K. et Iwasaki H. (1996). Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*.

Turenne N. (2000). *Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles.* Thèse de doctorat, Université Louis Pasteur, Strasbourg.

Citations Titles Standardization Using Information Retrieval Techniques

Rogério Mugnaini, Esteban Fernandez Tuesta, Adalberto Otranto Tardelli

OMS/OPS/BIREME
Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde.
Rua Botucatu, 862 - 04023-901 São Paulo – Brasil
rogerio@bireme.br

Abstract

Citation analysis is an impossible method in many bibliographic databases. Few peripherical countries have worked with this type of data, using for this purpose, the information provided by ISI. The Brazilian database from SciELO Project presents near 475 thousands citations to scientific journals articles. Considering that citations demand several standardization work on titles, this paper presents a methodology, based on information retrieval, for standardization though ISSN database search. The result is the standardization of approximately 85% of the titles.

Keywords: Bibliometrics, Textual Statistics, Information Retrieval, Citation, Similarity, Brazilian, SciELO Project.

1. Introduction

Many questions related to the utilization of the ISI's (Institute for Scientific Information) databases for scientific production (? output) study of peripheral countries and also for European countries, above all the non-English-speaking, can be found in the literature, in the first place because the 'science' of the first ones cannot be considered mainstream.

Two postulates are in the basis of the bibliometrical analysis techniques (Rostaing, 1996): first, "the publication is, in a scientific context, a representation of the scientific activity of its author;" second, "exists relationship among the ideas of "the proper author and the ideas acquired from the work of its peers."

The citation analysis is a bibliometrical technique that makes sense, according to those postulates, within Merton's paradigm, in that the development of science is given according to own internal logic, objectifying to generate new knowledge; science whose product is reflected in written scientific communication, particularly in the scientific journals.

The impact factor has been quite questioned within the reasons raised above. The few journals indicated in the ISI's databases of countries as Brazil, have no sufficient visibility with regard to the Americans, for example. Certainly the impact of those journals, and of the large majority whose visibility is restricted to the national libraries' shelves are restricted to the local journals.

The citation analysis becomes increasingly important, as it is adopted as quality indicator of scientific journals. Being utilized for the quantification of the impact factor, measure created by the ISI since the 1960s (Garfield, 1994) is widespread among the science and technology's systems of many countries of the world.

Within the bibliographic databases owners, a minority has this type of data, whose analysis always requires a severe standardization treatment. This work aims to present a methodology to obtain a satisfactory level of correction and standardization, utilizing a information retrieval based process (Tardelli, 2003)[1], that is given through previous treatment of the cited titles and through a standardization database.

2. SciELO Project[2]

SciELO (Scientific Electronic Library Online) project is the result of the efforts of BIREME (Latin American and Caribbean Health Sciences Information Center) and of the FAPESP (State of São Paulo Research Foundation), and envisages the development of a common methodology for the preparation, storage, dissemination and evaluation of scientific literature in electronic format. It is contains currently a total of 114 journals, summing approximately 32 thousand articles, and around 475 thousand citations to journal articles, whose titles, tabulated, sum approximately 70 thousand variations.

It has their databases built in CDS/ISIS[3] (Computerized Information Service/ Integrated Scientific Information System) standard, which permits the establishment of relationships between register of a database or between records of different databases. Operations are done with CISIS[4] utilities, programs that permit the administration, manipulation and realization of operations (many of them nonexistent in the conventional versions of Microisis - CDS/ISIS 3.07 and WinISIS 1.31) between several CDS/ISIS databases.

With the ISSN database in the same standard, which can serve as reference for standardization of the cited titles, some procedures are sequentially adopted for the constitution of a titles standardization database. This standardization work becomes necessary due to the lack of uniformity in the writing of the titles of cited journals in the bibliographic reference set of the indexed articles in the database.

3. Process of standardization

A flow chart of the standardization process can be observed in the following figure.

Firstly, the title is sought in the form that was cited (V1), in the standardization database (if found, Y1). Not finding, this title will be sought directly in the ISSN basis (if found, Y2) and if it still was not found will be submitted to small substitutions of point and space (V1*) and sought in the basis ISSN (if found, Y3).

The errors of writing (digitations) are the leading cause of the non-uniformity of cited titles; in addition, abridgements are very common, since many titles are known by the abbreviated title; despite this, the most frequent citations have written titles (complete or abbreviated) correctly. The correction of titles keyed in incorrectly can be given utilizing a tool (Trigrams) that associates and orders expressions per the degree of similarity among them. Thus, the title cited incorrectly, with smaller frequency, can be replaced by a similar written form (V2), however more utilized (if it had), since this last tends to be more correct. Then this title will be submitted to the basis of standardization (if found, Y4) and if not found, continues the searches in the basis ISSN (Y5 and Y6), just like in stages Y2 and Y3.

[1] Available at: < http://www.crics.info/reuniao_bvs3/program/docs/es/1 >. Access in: August, 5 2003.
[2] Available at: < http://www.scielo.org >. Access in: August, 5 2003.
[3] Available at: < http://www.unesco.org/webworld/isis >. Access in: August, 5 2003.
[4] Available at: < http://productos.bvsalud.org/html/pt/home.html >. Access in: August, 5 2003.

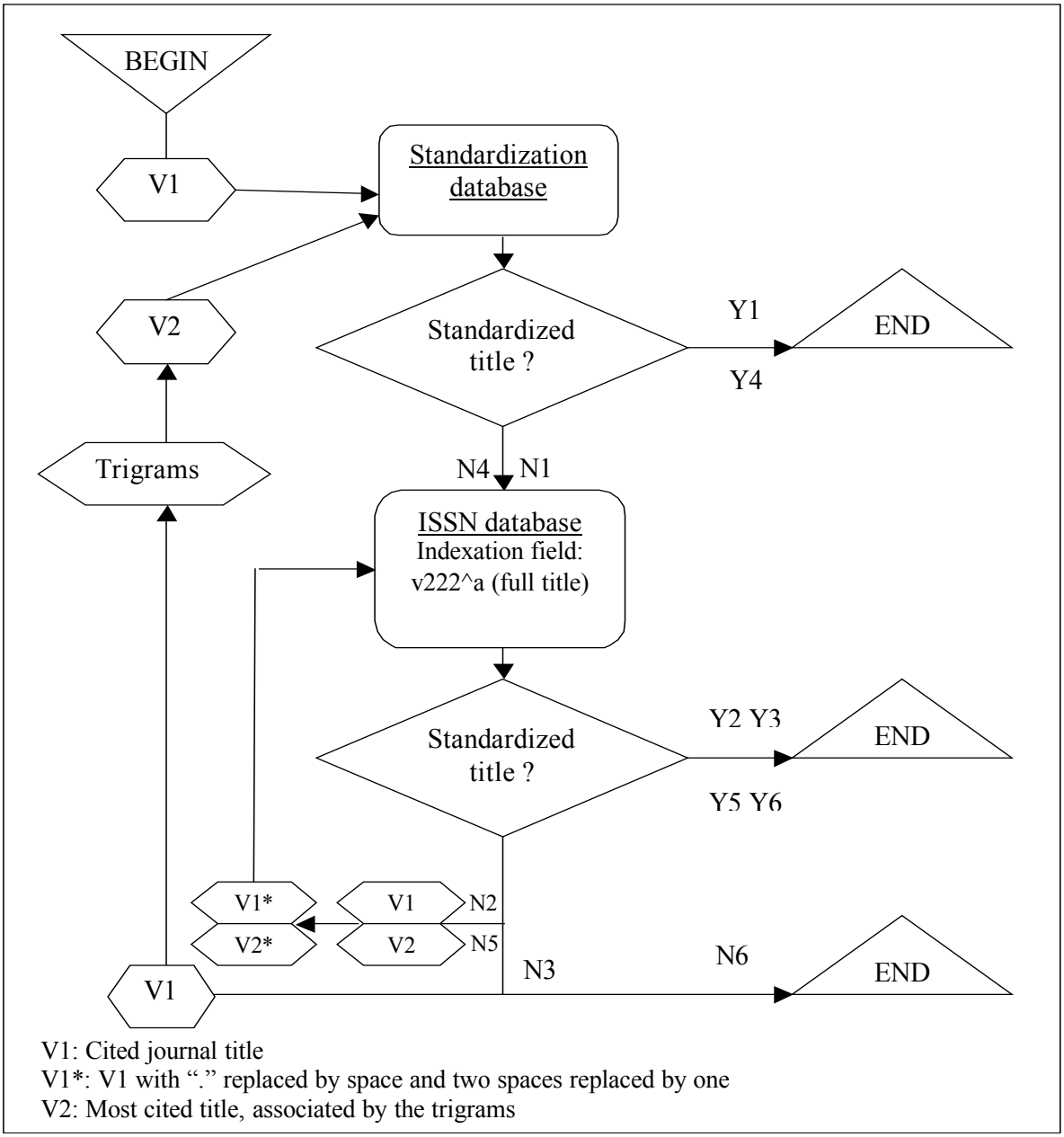


Figure 1. Flow chart of the standardization process

A description of the stages of the standardization process is presented in the above subsections.

3.1. *Standardization database*

3.1.1. *First stage*

The first stage consists of a search of full titles in the ISSN database. That search was defined in order to retrieve, utilizing the title of the journal cited in SciELO database, the corresponding title in the ISSN database. To this end, were defined firstly, the fields of the ISSN database that would turn "investigable", through a Fields Selection Table - FST.

A database in the CDS/ISIS standard contains a set of files that composes it, within which, the Inverted File. This file contains all the terms that can be used as points of access during the retriever of the data, and for each term, a reference list of the records of a Master File from

which the term was extracted. The collection of all the access points for the database is called of "Dictionary". The Inverted File can be thought as a remissive index of the Master File, where are defined the investigable elements of the database, through the Fields Selection Table, that contains the fields to be inverted, and the technique of indexing to be used for every each. There are 5 indexing techniques identified by numerical codes: ø to 4. In this stage it was utilized the technique ø, that constructs an element of search based on every line extracted from the format (information inside one or more fields, with optional adding of any string). Used to index entire fields or sub fields with up to 60 characters.

The fields of the ISSN database selected for realization of the search were:

210		**Abbreviated key title**
	&a	Abbreviated key title
	&b	Abbreviated qualifying information distinguishing otherwise identical key titles
	&c	Abbreviated qualifying information distinguishing otherwise identical abbreviated key titles
222		**Key title**
	&a	Key title
	&b	Added qualifying information distinguishing otherwise identical key titles
245		**Title proper**
	&a	Title proper or common title part of the title proper
	&s	Section, sub series, supplement designation
	&u	Section, sub series, supplement title
246		**Variant title(s)**
		Access to portions of titles and developed forms of key title
		Parallel titles
		Other forms of title not specified
		Cover title
	&a	Title

Source: http://www.issn.org:8080/English/pub/tools/format

The content of each of those fields was utilized combining the contents of its sub fields and making substitutions of point by space, and two spaces by one. In that way, a record of the ISSN basis whose search fields are presented as follows:

```
210  "  ^aTherapia^bHels.^cSuom. p."
222  " 0^aTherapia^bHelsinki"
245  " 0^aTherapia"
246  " 3^aAstra therapia"
246  " 3^aAstraZeneca therapia"
```

It will be represented by the terms of access as follows:

```
THERAPIA
THERAPIA HELS
THERAPIA SUOM P
THERAPIA HELS SUOM P
THERAPIA HELSINKI
ASTRA THERAPIA
ASTRAZENECA THERAPIA
```

The search expression, at this stage, utilizes the cited title, complete, that can be associated to more than one record of the ISSN database. At this time single relationships between cited title and title of ISSN database are barely retrieved.

The citation with journal title "THERAPIA", matches with the string "Therapia" (ISSN database) of Helsinki, finds, in the basis ISSN also the editions, with equal title, of Buenos Aires, Barcelona and Bratislava. It is concluded that an incomplete citation is with difficulty identified, and as much fewer words in the title, greater is the chance of ambiguity.

The resulting database presents a total of 17,175 single relations, between cited title and title in the ISSN, representing approximately 24.3% of the different cited titles, which covers a total of 194,866 citations (41.1%).

3.1.2. *Second stage*

In the second stage the search is carried out utilizing the words of the cited title, instead of the complete expression. This type of indexing can be obtained with technique 4, that constructs an search element based on each sub field, considering as search element any sequence of alphanumeric characters. The Fields Selection Table of ISSN database in that stage, barely utilizes the sub field "a" of field 222, indexed word-to-word.

The citation made to the abbreviated title "CAN J ANIM SCI", will be used as expression of search, having each one of the four words any termination ("$"), and separated by Boolean operator "and", turning out: "CAN$ * J$ * ANIM$ * SCI$", that retrieves the title "Canadian journal of animal science" of the ISSN database.

The resulting database from this stage presents a total of 18,697 single relations, between cited title and title in the ISSN. However, of those relations 10,033 are exclusive with the relations of the first stage, representing then, around 14.2% of the different cited titles, which covers a total of 24,203 citations (5.1%).

3.2. *Correction Process*

3.2.1. *Similarity indexes*

The search for relative similarity between documents, was utilized the article written by G. Salton and C. Buckley (1988). The words that compose a document, are extracted, and placed in order to form a vector,

$$D = (t_0, w(d,0); t_1, w(d,1); \ldots; t_t, w(d,t))$$

where each t_i represents a word that composes the D document and each $w(d,i)$ represents the weight of the word t_i in the D document. Similarly, the required information, or "Query" is represented in vector form as

$$Q = (q_0, w(q,0); q_1, w(q,1); \ldots; q_t, w(q,t))$$

where each q_i it represents a word of the "query" Q and each $w(q,i)$ represents the weight of that word in the "query".

When a standardized size is used, we obtain the well-known similarity formulae

$$sim(Q,D) = \frac{\sum_{k=1}^{t} w(q,k) * w(d,k)}{\sqrt{\sum_{k=1}^{t} w(q,k)^2} * \sqrt{\sum_{k=1}^{t} w(d,k)^2}}$$

Within the objectives of the "National Library of Medicine", of automatic methods of indexing "indexing initiative", the work written by Aronson *et al.* (1999) presents a section devoted to "trigrams". This method consists of the construction of vectors, not barely as being words, but, sets of three anagrams of each word, establishing weights for each "trigram" depending on its location in the word.

The results of the similarity between Q and D will remain sorted descendant, in such a way that among the first ones, will be those "queries" more similar to the documents.

4. Results obtained

An amount of approximately 85% of the titles was standardized, being distributed as the following table. The different moments of the process of standardization share individually for the conversion of the titles, being able to observe which of them are the more significant.

Stages	Frequency	Percentage
Y1	214,888	*46.6*
Y2	171,803	*37.3*
Y3	2,403	*0.5*
Y4	7,798	*1.7*
Y5	6,434	*1.4*
Y6	68	*0.0*
No standardized	57,586	*12.5*
	460,980	***100.0***

Y1 and Y4b are the stages where the titles found are perfectly identified in the ISSN database, amounting to a total of approximately 50% of the citations. In the other four stages they are already retrieved more than one title, not being able to be aimed with precision the title cited for lack of information of the citing author.

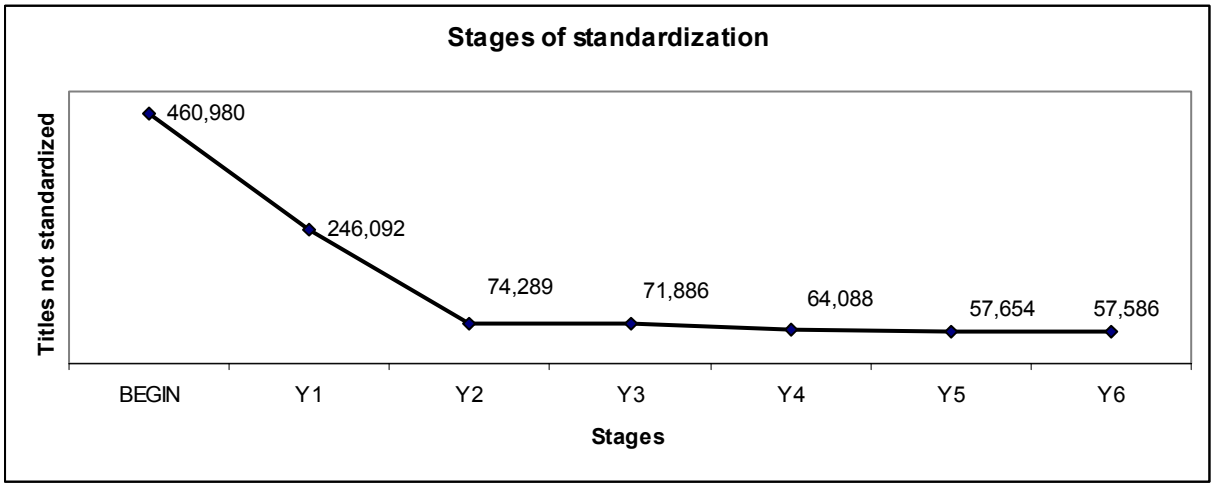


Figure 2. Stages of standardization

The first two stages have principal role in all the process of standardization, since they represent almost 96% of the standardized total.

References

Aronson A.R., Bodenreider O., Chang H.F, Humphrey S.M., Mork J.G, Nelson S.J., Rindflesch T.C. and Wilbur W.J. (1999). *The indexing initiative*. A report to the Borrad of Scientific Counselors of the Lister Hill Nacional Center for Biomedical Communications.

Aronson A.R., Bodenreider O., Chang H.F., Humphrey S.M., Mork J.G., Nelson S.J., Rindflesch T.C. and Wilbur W.J. (2000). The NLM indexing initiative. In *2000 AMIA Annual Fall Symposium*: 17-21.

Garfield E. (1995). Quantitative analysis of the scientific literature and its implications for science policymaking in Latin America and the Caribbean. *Bulletin of PAHO*, vol. (29): 87-95.

Humphrey S.M., Rindflesch T.C. and Aronson A.R. (2000). *Automatic Indexing by discipline and high-level categories: methodology and potential applications.*

Rostaing H. (1996). *La bibliométrie et ses techniques*. Sciences de la Société.

Salton G. and Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Procesing & Management*, vol. (24): 513-523.

Tardelli A.O. (2003). Indización automática y "vector mining": herramientas para recuperación y vinculación de información en las fuentes de información de la BVS. In *3a Reunión de Coordinación Regional de la BVS*.

ttda - une librairie R pour l'analyse de données textuelles

Jean-Pierre Müller

Université de Lausanne – Faculté des Sciences Sociales et Politiques – BFSH #2
1015 Lausanne – Suisse
jean-pierre.mueller@dssp.unit.ch

Abstract

This contribution shows how it is possible to use "GPL" or "free" software to realize statistical analysis of textual data. R has been used as a base, as it becomes a software reference in the statistical community. A free solution was also found for lemmatization.

Résumé

Comment utiliser des logiciels avec des licences de type « GNU Public Licence » ou « libres » pour réaliser des analyses de données textuelles ? R devenant un logiciel statistique de référence, c'est naturellement qu'il a été utilisé comme élément intégrateur par la création d'un module spécifique pour le traitement de données textuelles. Une solution a été aussi trouvée pour la lemmatisation des formes. Les traitements statistiques proprement dits sont alors réalisés avec des modules connus de R.

Mots-clés : R, ispell, ttda.

1. Introduction

1.1. *Motivation*

Les logiciels d'analyse de données textuelles sont payants, pour une majorité, et sont donc difficiles à mettre à disposition des étudiants pour des travaux pratiques ou des mémoires de fin d'études. Par ailleurs, les étudiants et chercheurs utilisent de nombreux systèmes d'exploitation ; hors pour certains de ces derniers, il est extrêmement difficile de trouver une solution pour le traitement statistique de textes.

Cette contribution évoque les différents éléments ayant permis de développer une librairie de routines pour le logiciel R, d'abord pour MacOS X, puis pour les systèmes linux.

1.2. *Autres logiciels*

Il existe de nombreux logiciels qui permettent de traiter des données textuelles : SPAD-T (édité par DECISIA), Lexico (réalisé par l'équipe universitaire SYLED-CLA2T), Alceste (édité par IMAGE), Tropes (édité par Acetic), Hyperbase (Institut de Linguistique Française, UFR Lettres, Arts et Sciences Humaines, Nice), Sphinx (Le Sphinx Développement). Souvent dotés d'une interface agréable, ils autorisent de nombreux traitements et introduisent des développements spécifiques (lemmatisation, méthodes factorielles et classifications, représentations diverses du corpus). Si ces logiciels sont tous disponibles pour la plateforme Windows-Intel, seuls trois fonctionnent sous Mac OS 9 (Lexico, Alceste et Hyperbase), aucun nativement sous Mac OS X (bien qu'une version d'Alceste soit annoncée) ou Linux.

1.3. *Cahier des charges*

Le cahier des charges était relativement sommaire :

- Implémenter ou avoir accès à une partie des techniques décrites dans un ouvrage de référence comme celui de Lebart et Salem (1994).
- Traiter un corpus se présentant sous la forme d'un fichier unique, les uci (unités de contexte initial ou parties) étant marquées de manière usuelle.
- Évaluer la solution choisie et ses possibilités d'évolution.

Les choix d'implémentation ont été les suivants :

- La balise `<uci=nombre>` est utilisée pour identifier le début d'une uci dans le corpus.
- La balise `<par>` peut être utilisée pour identifier le début d'un paragraphe.
- Les mots débutant par une étoile `*` (mots étoilés) sont considérés comme des marques de variables supplémentaires. Ils peuvent être placés partout dans l'uci.
- Les accolades peuvent être utilisées sur la même ligne pour inclure des commentaires dans le corpus : `{ commentaire }`.
- Les balises doivent être séparées par des espaces ; ainsi la chaîne `<uci=3><par>` ne sera pas reconnue comme un marquage valide.

2. R comme élément intégrateur

R (Ihaka et Gentleman, 1996) est désormais un logiciel largement connu dans la communauté statistique. Reprenant la syntaxe de S, il permet de réaliser de nombreuses analyses et d'écrire des procédures spécifiques pour certains traitements. Il est librement disponible pour de nombreuses plateformes (http://cran.r-project.org/).

Parmi les autres raisons prépondérantes ayant entraîné son choix comme élément intégrateur, il faut citer :

- La disponibilité de nombreux modules d'analyse de données (`ADE-4` ou `multiv`, par exemple) pour le traitement des tableaux lexicaux.
- `grep` est intégré à R, ce qui permet des traitements efficaces des chaînes de caractères.
- Bien que R soit interprété, de nombreuses fonctions sont implémentées en C ou en Fortran, ce qui lui permet d'avoir une bonne efficacité générale.

2.1. *Lecture du corpus*

La lecture du fichier du corpus est réalisée à l'aide de la fonction :

```
ttda.get.text(textfile = "")
```

Si le paramètre `textfile` est vide, un dialogue de sélection de fichier est présenté à l'écran. Cette fonction retourne un vecteur type (chaîne de) `character` contenant les lignes du fichier où une balise `<line>` a été ajouté au début de chaque ligne. Cette fonction utilise la fonction native `readLines` de R.

2.2. *Segmentation*

La segmentation du corpus en occurrences est réalisée par la fonction :

```
ttda.segmentation(textlines,
   separators = ttda.util.separators(),
   with.hyphen = TRUE,
   upper8 = ttda.util.get.modifier()[1],
   lower8 = ttda.util.get.modifier()[2])
```

Elle retourne un vecteur de type (chaîne de) `character` contenant les occurrences et les balises du corpus. Les séparateurs utilisés par défaut sont ceux retournés par `ttda.util.separators`, dont le résultat dépend du système d'exploitation de l'ordinateur utilisé. Le paramètre `with.hyphen` précise comment traiter les tirets. Si `TRUE` est utilisé, toute suite de tirets, précédée ou suivie d'espaces, sera considérée comme caractères séparateurs. Cette fonction remplace les caractères majuscules par les minuscules correspondantes. Le remplacement des caractères majuscules accentués en minuscules accentuées est réalisé par les paramètres `upper8` et `lower8`. Par défaut, les résultats de la fonction `ttda.util.get.modifier`, qui donne la table de conversion selon l'OS, sont utilisés pour cette transformation.

La fonction native `strsplit` est utilisée pour réaliser la segmentation et la fonction `chartr` pour le passage des majuscules aux minuscules.

2.3. *Création du dataframe du corpus*

Ce tableau de données contient un enregistrement par occurrence du corpus. Cinq variables décrivent l'occurrence :

`uci` : le numéro de l'uci qui contient l'occurrence.

`uce` : le numéro de l'uce (partitions des uci en phrases) qui contient l'occurrence ou 0 si elles n'ont pas été calculées par `ttda.uce`.

`par` : Le paragraphe qui contient l'occurrence, 0 si ils n'ont pas été indiqués dans le corpus.

`line` : la ligne qui contient l'occurrence.

`graphical.forms` : l'occurrence, comme valeur d'un vecteur de type `factor`. Les `levels` de cette variable constituent un dictionnaire des formes du corpus.

Cette transformation est réalisée par la fonction :

```
ttda.forms.frame(occurences, remove.keys = TRUE)
```

à partir du résultat de la fonction `ttda.segmentation`. Le paramètre `remove.keys` précise si les balises doivent être conservées comme occurrences dans le tableau de données.

A défaut d'avoir trouvé une méthode plus élégante, cette transformation utilise des boucles sur les positions des balises.

3. Lemmatisation

La lemmatisation du texte est réalisée par `ISPELL`. `ISPELL` est un dictionnaire orthographique disponible pour les plateformes *nix et linux (attention : dans certaines distributions linux, `ISPELL` n'est qu'un script qui appelle le correcteur orthographique `aspell`). Une implémentation pour MS Windows existe, mais n'a pas été testée. `ISPELL` est couvert par la licence de l'Université de Berkeley, il est disponible sur Internet à l'adresse : http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html. Il commence par comparer le mot à corriger aux entrées de son dictionnaire, puis, si le mot n'y apparaît pas, combine celles-ci

avec une liste de préfixes et de suffixes. De nombreux dictionnaires et listes d'affixes sont disponibles pour les langues européennes. Une description très complète du fonctionnement d'ISPELL est donnée par Pythoud (1996).

Lemmatisation par ISPELL

`ISPELL` peut être mis sur un mode de fonctionnement particulier à l'aide du drapeau `-a`, ce qui permet de l'utiliser depuis d'autres programmes. Dans ce mode, pour chaque mot soumis, `ISPELL` retourne une ligne ayant l'un des formats suivants :

`* :` si le mot soumis a été trouvé dans le dictionnaire ; `+ <lemme> :` si le mot soumis peut être déduit de <lemme> en utilisant les règles d'affixation ; ou des lignes plus complexes débutant par les caractères `-`, `&`, `?`, `#`, si un mot composé, un ou des mots proches ou pouvant être déduits illégalement du dictionnaire ont été trouvés ou si rien n´a été trouvé.

La fonction :

```
ttda.lemmatisation(forms, ...)
```

réalise la lemmatisation en appelant `ISPELL` et en traitant de manière adaptée le résultat retourné par le correcteur orthographique. Il est conseillé de la placer dans une commande `transform` qui agit directement sur le dataframe.

Les paramètres suivants de `ttda.lemmatisation` précisent l'utilisation d'`ISPELL` sur la machine utilisée :

`ispell.enc` est l'encodage des caractères utilisé par `ISPELL`. Utiliser `ISOLatin1` sur Macintosh. La valeur par défaut `native.enc` devrait fonctionner pour les autres OS.

`dict` spécifie le dictionnaire à utiliser. Si le paramètre n'est pas renseigné, l'anglais est utilisé. Donner la valeur `francais` pour utiliser le dictionnaire français. Pour d'autres dictionnaires, consulter le site web de `ISPELL`.

`ispell.path` Le chemin d'accès à `ISPELL`. La valeur par defaut `/sw/bin/ispell` est valable pour l'installation faite par `fink` de `ISPELL` sous MacOS X.

4. Procédures spécifiques à l'analyse de données textuelles

4.1. *Fonctions utilitaires*

Il est conseillé de les effectuer dans une commande `transform` qui agit directement sur le dataframe.

La réduction à des formes non accentuées est réalisée par la fonction :
`ttda.remove.modifier(forms)`.

`ttda.match(forms, old.levels, new.levels)` vous permet de re-coder des formes manuellement.

4.2. *Outils d'analyse du dataframe du corpus*

La fréquence des formes peut être obtenue à l'aide de la commande R : `summary()`.

La fonction `ttda.hapax(forms)` permet d'obtenir la liste des hapax.

La fonction `ttda.pareto()` permet d'obtenir des graphiques des proportions des formes (ou occurences) selon leurs fréquences.

La fonction

```
ttda.concordances(forms, theform, uci, indices = NULL,
     n.before = 3, n.after = 3, sort = "text", out = "vector")
```

permet d'obtenir les concordances des formes spécifiées en paramètre.

4.3. *Création du tableau lexical*

La fonction `ttda.TLE(uc, forms, uc.names = NULL)` permet de construire un tableau lexical à partir du dataframe du corpus. Le paramètre `uc.names` permet de nommer les parties dans le tableau retourné.

4.4. *Outils d'analyse du tableau lexical*

Les deux fonctions :

```
ttda.get.suplvar(tle)
ttda.remove.suplvar(tle)
```

agissent sur le tableau lexical. La première permet d'obtenir une copie des colonnes du tableau lexical provenant de mots étoilés (variables illustratives). La seconde permet de supprimer ces mêmes colonnes du tableau lexical.

Les trois fonctions

```
ttda.f.speci.exact(),ttda.speci.pos(),ttda.speci.neg()
```

permettent d'étudier les formes spécifiques d'une partie, en particulier d'une uc. La première fonction permet de calculer la probabilité d'avoir un effectif supérieur ou inférieur à celui observé, les deux dernières d'extraire du tableau des probabilités les formes de spécificité négative ou positive.

R fournit tous les outils nécessaires pour faire des AFC, ACM, et autres techniques statistiques, en particulier les packages : `ade4`, `multiv`, `vegan`, `tree`, `rpart`, `mva`, `class`, `cluster`.

5. Résultats et limitations

La qualité de la lemmatisation dépend très fortement du dictionnaire utilisé ou simplement disponible. Le dictionnaire et le fichier d'affixes de `Francais-GUTenberg`, également développés par Pythoud (1999) donnent de bons résultats sur des textes français tout-venant. La lemmatisation réalisée vous surprendra peut-être, les lemmes étant parfois de forme féminine, `directeurs` étant lemmatisé `directrice`.

Deux problèmes sont toutefois posés par `Francais-GUTenberg` :

1) La lemmatisation des verbes n'est pas parfaite, car il utilise au moins cinq lemmes par verbe : les premières personnes du singulier du présent, de l'imparfait, du passé simple et du futur de l'indicatif, ainsi que l'infinitif. La solution pourrait être la création d'un fichier de congruences de ces lemmes.

2) Le problème de l'élision n'est pas réglé de manière satisfaisante dans la version actuelle de la librairie. Les formes élidées ne sont pas reconstituées par `ISPELL` si l'apostrophe est traitée comme caractère séparateur dans `ttda.segmentation` et, sinon, la forme élidée est perdue ou non traitée par `ISPELL` (par exemple, `s'attendre` est lemmatisé `attendre`,

`l'âme` comme `âme`, mais `j'assaillisse` est considéré comme un lemme valide)[1].

La taille du corpus pouvant être traité dépend de la mémoire disponible, toutes les opérations étant réalisées en mémoire vive. De nombreuses méthodes ne sont pas encore intégrées: segments répétés et unités caractéristiques, par exemple.

La relative simplicité de la programmation dans R permet d'envisager l'adaptation rapide de ces techniques dans `ttda`. L'existence de librairies permettant d'accéder à des données stockées dans des bases de données SQL permet d'envisager des procédures de traitement de corpus importants et les fonctions d'accès à Internet d'utiliser des données extraites de pages HTTP.

6. Disponibilité de ttda et exemples

La librairie `ttda` est disponible à l'adresse http://wwwpeople.unil.ch/jean-pierre.mueller/. Deux corpus sont fournis avec la librairie. `anmat.txt` est constitué de 330 annonces matrimoniales parues entre 1960 et 1990. `fed_paper.txt` contient les 85 essais dus à Alexander Hamilton, John Jay, ou James Madison parus d'octobre 1787 à mai 1788, dont le but était de convaincre les habitants de New-York d'accepter la future constitution américaine. Ils sont connus sous le nom de « The Federalist papers ». Des exemples complets de traitement de ces fichiers de données sont donnés dans le manuel français de la librairie.

Références

Ihaka R. et Gentleman R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, vol. (5/3) : 299–314.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Pythoud C. (1996). *Problèmes de la correction automatique de l'orthographe lexicale du français à travers une étude de cas: le correcteur orthographique Ispell et le dictionnaire Français IREQ.* Mémoire de licence, Section de linguistique, Université de Lausanne, http://www2.unil.ch/ling/cp/Francais-GUTenberg-v1.0.tar.gz

Pythoud C. (1999). *Francais-GUTenberg v1.0 - Manuel de référence*. http://www2.unil.ch/ling/cp/reference.pdf

[1] Entre deux maux, vous êtes donc invités à choisir celui qui vous gêne le moins !

Analyse automatique d'un discours spécialisé au moyen de grammaires locales

Takuya Nakamura

IGM – Université de Marne-la-Vallée
77454 Marne-la-Vallée Cedex 2 – France
nakamura@univ-mlv.fr

Abstract

A collection of specialized texts such as stock exchange market reports is formally analyzable into a set of sentence schemata. In this study, I take one example of these schemata, based on a specialized lexicon-grammar to describe formally a corpus constituted of an accumulation of stock exchange market reports found in a French daily newspaper. This linguistic description is also represented in the form of local grammars, in order to analyze automatically the corpus. The local grammars which represent one schema of sentence give rise to a non ambiguous analysis of approximately 22 % of the sentences in the corpus, which correspond to the totality of the sentences characterized by a sentence schema.

Résumé

Une collection de textes spécialisés, comme les bulletins boursiers, est susceptible d'être analysée formellement en une suite de schémas de phrase. Dans cette étude, nous avons pris un exemple de ces schémas de phrase, pour décrire notre corpus constitué de l'accumulation de rubriques boursières d'un quotidien français. Cette description linguistique est représentée également sous formes de grammaires locales, pour analyser automatiquement le corpus. Les grammaires locales, représentant un seul schéma de phrase ainsi décrit, ont reconnu, sans ambiguïté, 22 % des phrases du corpus, ce qui correspond à la totalité des phrases correspondant au schéma.

Mots-clés : analyse automatique d'un discours spécialisé, corpus économique, lexique-grammaire, grammaire locales.

1. Introduction

Les brefs textes journalistiques publiés régulièrement pour rapporter des événements qui se produisent dans un domaine particulier, comme les bulletins météorologiques ou les résultats sportifs, sont des discours spécialisés.

Les rapports boursiers, comme on en trouve dans un quotidien le lendemain d'un jour ouvré, en constituent un exemple. Nous allons montrer que l'impression de stéréotype que donnent ces rapports peut être caractérisée linguistiquement, en réduisant une partie significative des phrases du discours à un schéma de phrase, constitué de plusieurs composants formels. La méthode d'analyse linguistique du corpus est un lexique-grammaire spécialisé pour ce type de discours. La réduction des phrases du discours à un schéma de phrase est inspirée de Harris (1952 et 1963).

Les résultats de cette description linguistique sont manuellement transposés dans des grammaires locales (Gross, 1996 et 1997), pour pouvoir être appliquées au corpus. Grâce à une description linguistique minutieuse des phrases, ces grammaires sont lexicalisées en détail, à tel point que le recours à des noms de catégorie comme éléments de la grammaire locale, qui produit souvent des résultats erronés, peut être évité. Les composants des grammaires locales

sont clairement organisés après analyse syntactico-sémantique des phrases, afin que leur application donne des résultats sans ambiguïté. Cette double méthode - description linguistique et construction de grammaires locales - est libre des contraintes techniques externes qui sont toujours imposées lorsqu'on fait appel à un formalisme particulier de représentation informatique. Les résultats de la reconnaissance automatique sont extrêmement lisibles et explicites et ils peuvent être utilisés à des fins diverses, comme la traduction automatique ou la recherche d'informations dans les corpus.

Cette étude montre l'utilité d'une description linguistique des phrases du corpus préalable à la construction d'un système d'analyse automatique. Si l'on formalise un certain nombre de phrases réelles qui apparaissent dans le texte au moyen du lexique-grammaire, la construction des grammaires locales pour la reconnaissance de ces objets en est largement facilitée. D'autre part, le schéma de phrase étant constitué en tenant compte de tous les composants syntactico-sémantique nécessaires pour caractériser des informations spécifiques au discours y compris les compléments, considérés, en dehors de ce discours particulier, comme circonstantiels, donc aléatoires, la part d'ombre (cf. des insertions imprévisibles laissées dans les grammaires locales) est considérablement diminuée.

Dans ce qui suit, après la présentation du corpus étudié dans la section 2, nous montrons les méthodes linguistiques de l'analyse du corpus dans la section 3, puis la façon de confectionner les grammaires locales basée sur les descriptions linguistiques dans la section 4.

2. Corpus

Le corpus est constitué de rubriques *Valeurs France* parues dans le quotidien *Le Monde*. Dans cette rubrique quotidienne, on trouve de 3 à 6 brefs commentaires sur les mouvements de certaines actions échangées à la Bourse de Paris et les activités économiques ou les phénomènes sociaux liés à ces mouvements. Chaque commentaire comporte environ deux ou trois phrases.

La période de collecte va du 18 janvier 2001 au 24 août 2001, soit au total 146 rubriques comportant à peu près 560 commentaires. La taille du corpus est de 175 kilo octets.

3. L'analyse du corpus

3.1. *L'analyse du discours*

Selon Harris (1952), un discours est réductible à une succession de séquences de plusieurs catégories formelles. Ces catégories sont des classes d'équivalence, établies d'un point de vue distributionnel et transformationnel, en fonction du discours étudié. Nous considérons que l'analyse harrissienne du discours n'est pas applicable telle quelle à un commentaire trouvé dans le corpus qu'on étudie ici, puisque le nombre de phrases trouvées est petit et que les phrases contenues dans le discours sont disparates.

Si on prend pour un seul discours une accumulation de commentaires boursiers, il en émerge cependant une séquence de plusieurs catégories syntactico-sémantique définies pour le corpus boursier, désormais appelée schéma de phrase, qui représente des occurrences concrètes de phrases. D'après les résultats obtenus au terme de notre étude, un quart du corpus correspond à un seul schéma de phrase, ce qui n'est pas loin de la tentative de Harris, qui a essayé de réduire un texte entier à une succession de plusieurs séquences de catégories.

La direction donnée par Harris dans le domaine de l'analyse du discours a été poursuivie au fur et à mesure que se développe l'analyse automatique de textes d'un domaine spécifique, au moyen du formalisme de la grammaire en chaîne. Les travaux sur les textes médicaux (Hirschman et Sager, 1980) et les recherches sur les rapports boursiers ou les bulletins météo

(Kittredge, 1980) ont été effectués dans ce cadre. Ces dernières traitent du même type du phrase qu'on étudie dans cet article mais l'accent a été mis sur la façon de caractériser, de point de vue des sous-langages (*sublanguage* en anglais), les phrases dans le discours d'un domaine spécifique.

3.2. *Deux types d'informations dans le corpus*

Dans les rapports boursiers, on trouve souvent deux types d'informations, liées par une relation de cause à effet. Le premier type se trouve dans des phrases qui décrivent les mouvements en valeur d'actions particulières, observés à la Bourse. Cette information est donc spécifique à la Bourse. L'autre type relate un phénomène économique ou social qui pourrait être la cause de ces mouvements. Cette information peut être extra-boursière. En voici deux exemples types :

(1) *A 34 euros, l'action X a augmenté de 0,25 %, mercredi 26 mai, à l'ouverture de la Bourse de Paris. X a annoncé la veille le licenciement massif de trois quarts de ses employés.*

(2) *L'action X a débuté la séance du mercredi 26 mai en baisse de 1,1 %. La production du café en Amérique du Sud a subi de gros dégâts à cause d'une chaleur anormale.*

Le discours imaginé (1) est constitué de deux phrases. La première décrit le mouvement en valeur des actions d'une entreprise nommée *X* et la deuxième une activité de cette entreprise, qui est relatée ici parce qu'elle est censée avoir une influence sur la chute de valeur de l'action décrite dans la première phrase. Dans le discours également imaginé (2), la première phrase présente une similitude de forme avec celle du discours (1) et la deuxième, qui décrit un phénomène naturel, ne peut pas être comprise dans ce contexte si le lecteur n'a pas de connaissance préalable de la société *X*, qui est, par exemple, un exportateur du café.

Comme nous allons le voir, le premier type de phrase dans les deux discours peut être décrit formellement au moyen du lexique-grammaire. Mais le second type de phrase est plus difficile à schématiser formellement, parce que la phrase décrit non pas des activités boursières mais divers phénomènes dans le monde réel, impliquant divers prédicats et divers arguments.

Il existe des phrases qui appartiennent au deuxième type de phrase donné ci-dessus et qui décrivent les activités financières et commerciales d'une société comme suit :

(3) *X a (annoncé + publié + prévu) le chiffre d'affaires pour le second trimestre 2003 en (hausse + baisse) de 0,3 %, à 23 milliards d'euros.*

Elles sont susceptibles d'être décrites à partir du schéma de phrase qui sera expliqué plus loin pour analyser le premier type de phrase. Mais nous n'avons pas tenu compte de ces phrases dans cette étude.

Nous nous sommes limité à la description du premier type de phrases mentionné plus haut, (les premières phrases des discours (1) et (2)) et leurs variantes formelles.

La description syntaxique de ces phrases est basée sur le lexique-grammaire[1], qui recense pour un prédicat donné la nature et le nombre de ses compléments possibles et les propriétés syntaxiques et transformationnelles qu'elles présentent. Cette description se traduira sous forme d'une table syntaxique dont les entrées sont constituées par les prédicats et où les colonnes indiquent les propriétés distributionnelles et syntaxiques.

[1] Pour connaître le détail, voir Gross (1975 et 1981) et Leclère (2002).

3.3. *Les phrases exprimant la variation d'une valeur*

Le type de phrase traité décrit *la variation en valeur d'une action particulière cotée à la Bourse de Paris*. En voici quelques exemples types :

(4) *ABC gagnait 2,84 %, vendredi 24 août dans les premières transactions, à 18,10 euros.*
(5) *L'action ABC s'appréciait de 0,94 %, vendredi dans les premiers échanges, à 37,48 euros.*
(6) *Le titre d'ABC (a fait un bond + a bondi) de 3,29 %, à 16,78 euros, mercredi matin*
(7) *L'action du spécialiste de la carte à puce ABC restait stable dans les premières transactions, lundi 20 août, à 3,09 euros.*

L'information essentielle, dans les rapports boursiers, est syntaxiquement structurée autour de plusieurs dizaines de prédicats (verbaux, nominaux ou adjectivaux, i.e. les séquences soulignées dans les exemples ci-dessus) et leurs compléments. Diverses formes de phrases ainsi construites partagent toutes une unité de sens stable, qui est la variation de quantité. Le sens de variation est rendu syntaxiquement explicite par la présence de deux compléments numériques, l'un relatif et l'autre absolu.

Les sujets de ces prédicats sont réalisés par un ensemble de groupes nominaux qui sont principalement juxtaposés d'une manière appositive (cf. (5)-(7)) autour de noms têtes comme *action, titre* ou *valeur*. Les noms propres des sociétés dont les transferts de l'action sont rapportés dans les textes ont été répertoriés et jouent un rôle syntaxique de sujet par métonymie (cf.(4)).

Les deux compléments numériques ont un lien fort avec les verbes, malgré leur nature adverbiale. Surtout, le complément numérique relatif est indispensable pour définir un certain nombre de verbes qu'on appelle ici « verbes de variation ». Il entretient donc un lien plus étroit avec ces verbes que le complément numérique absolu.

D'autres compléments, cette fois clairement adverbiaux, de temps et de lieu, sont considérés comme des compléments spécifiques de phrase. Le complément de lieu est plus facilement omis, étant implicitement constant[2], que celui de temps. L'information du temps est évidemment cruciale dans un rapport boursier en tant qu'élément d'observation.

Par ailleurs, un même prédicat peut apparaître sous forme de variantes formelles. Plusieurs prédicats verbaux utilisés dans ce contexte peuvent subir une transformation de nominalisation au moyen de verbes supports (Gross, 1981). Ils se rangent dans le même paradigme syntaxique. L'exemple (6) montre un cas de nominalisation possible.

L'organisation linéaire des prédicats et de leurs compléments est assez régulière. La nature du corpus constitué de rapports factuels fait qu'il n'y a de phrases ni interrogatives, ni impératives. Toutes les phrases observées sont déclaratives.

Les prédicats peuvent être classés, sémantiquement, dans trois groupes : ceux qui expriment le mouvement en hausse d'une valeur, ceux qui exprime le mouvement contraire, c'est-à-dire la baisse d'une valeur, et le dernier type qui exprime la stabilité.

Le type de phrase étudié se schématise comme une suite de composants formels :

A V D% DN AT AL

Chaque composant doit être considéré comme une classe d'équivalence, à l'intérieur de laquelle toutes les variantes formelles de la classe doivent être énumérées. ***A*** regroupe des

[2] Dans notre corpus, le lieu d'observation est toujours à la Bourse de Paris.

groupes nominaux fonctionnant comme sujet des prédicats de variation, ***V*** les prédicats de variation, ***D%*** les compléments numériques relatifs et ***DN*** les compléments numériques absolus. ***AT*** représente les adverbes de temps et ***AL*** les adverbes de lieu.

4. La construction de grammaires locales

4.1. *Le système UNITEX et les grammaires locales*

Le système UNITEX (Paumier, 2003) est un système de traitement automatique de la langue naturelle qui permet d'étiqueter les textes avec les dictionnaires électroniques existants et/ou créés selon les besoins. Les grammaires locales sont des graphes d'automates à états finis[3] qui, partant du point de départ, reconnaissent les unités linguistiques qui se trouvent sur un chemin. Lorsque un chemin a été exploré et qu'on arrive au point final, la séquence a été reconnue (Gross, 1997).

Le système UNITEX permet d'appliquer les grammaires locales aux textes étiquetés lexicalement.

Nous analysons notre corpus au moyen d'un ensemble de grammaires locales (266 graphes) dont l'unité maximale correspond au schéma de phrase donné plus haut et d'un dictionnaire des noms propres créé spécialement pour analyser le corpus (416 entrées).

Certains des composants du schéma de phrase sont exposés ci-dessous, soit sous forme d'une grammaire locale, soit sous forme d'une expression rationnelle.

4.2. *Le composant* A

Le constituant *A* du schéma de phrase ci-dessus représente les groupes nominaux suivant :

(L'action + le titre + la valeur) ((de + E) Npr + de Ns Npr + E)[4]

Les opérations d'effacement opérées sur les séquences (8) permettent d'obtenir la classe d'équivalence de groupes nominaux suivante. Ils sont co-référents dans le contexte en question :

(9) *(l'action + le titre + la valeur) de [la société de construction automobile]*$_{Ns}$ *[ABC]*$_{Npr}$
= (l'action + le titre + la valeur) de [la société de construction automobile]$_{Ns}$
= (l'action + le titre + la valeur) (de + E) [ABC]$_{Npr}$

Les groupes nominaux (9) sont ultimement réductibles aux groupes nominaux suivants :

(10) *((l'action + le titre + la valeur) + ABC)* (= (9))

Ces séquences sont donc considérées comme produites à partir de la séquence la plus longue en y opérant les effacements. Ce genre de faits linguistiques est facilement exprimé au moyen d'expressions régulières comme (8) ou (9), ou de façon encore plus lisible, de grammaires locales (Figure 1). Les noms propres de sociétés sont recensés dans un dictionnaire électronique (Figure 2).

[3] Cette équivalence peut être vraie dans le domaine du traitement automatique de la langue naturelle.

[4] *Npr* désigne ici les noms propres. *Ns* désigne un ensemble de groupes nominaux comportant une sous-classification des activités. Ex : *société de production audiovisuelle, entreprise de fournitures de bureau*, etc. *E* représente l'élément vide. ''+'' représente un choix dans un paradigme.

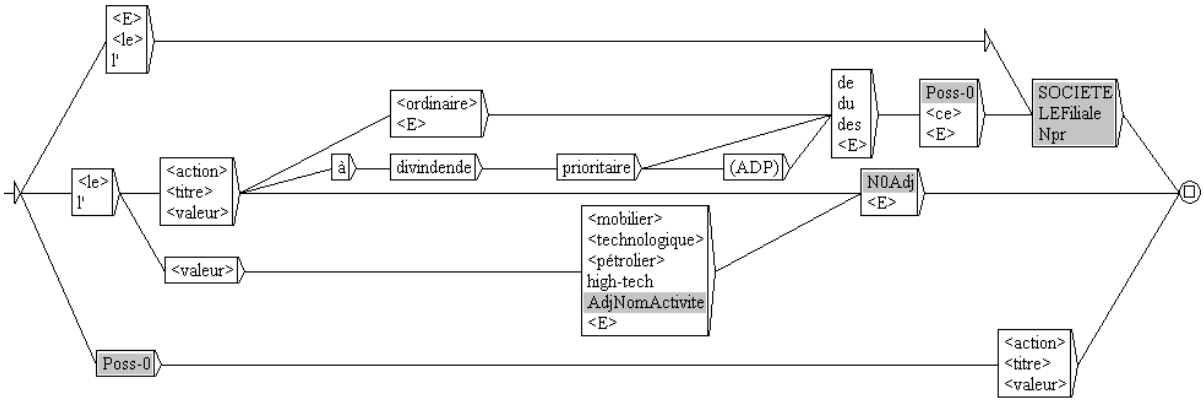


*Figure 1. Grammaire locale **A***
La flèche qui se trouve à gauche montre le point de départ et le carré à l'intérieur d'un cercle à droite est un point final. Les boîtes grises sont un appel à sous-graphe.

```
138 Eridiana Béghin-Say,.N+propre
139 Essilor,.N+propre
140 Etam Développement,.N+propre
141 Eurofins Scientific,.N+propre
142 Euronext,.N+propre
143 Euronext Paris,.N+propre
144 Europ@web,.N+propre
145 Eurosport France,.N+propre
146 Eurosport International,.N+propre
147 Eurotunnel,.N+propre
148 Executive Life,.N+propre
```

Figure 2. Dictionnaire de noms propres de sociétés

4.3. *Le composant V*

Le constituant *V* regroupe ici toutes les variantes formelles des prédicats de variation. La catégorie se subdivise en plusieurs classes. Si le prédicat est verbal, la transitivité par rapport au complément de mesure relative est un critère de subdivision. Si le prédicat est nominal, plusieurs sous-classes sont créés en fonction des verbes supports choisis.

Les verbes transitifs qui prennent comme objet direct le complément de mesure relative sont donnés ci-dessous. Cette propriété est marquée dans la colonne *NVDnum%* de la table (cf. Annexe). Ils forment la classe *Vvt* :

(11) *gagner, perdre, prendre, s'adjuger, céder, abandonner, regagner*

Avec les verbes intransitifs, le complément de mesure relative apparaît précédé de la préposition *de* :

(12) *progresser, chuter, baisser, dégringoler, bondir…*

Ces verbes forment la classe *Vvi* et la propriété figure sous la colonne *NVDeDnum%* dans la table.

Les prédicats peuvent être nominalisés. Selon les verbes supports, nous pouvons avoir dans la classe *V* les combinaisons de *Vsup* et *Npred* suivants :

(13) $[(accuser + enregistrer + connaître + ...)]_{Vsup}$ *UN (progression + baisse) de D%*

(14) *[(pointer + être + s'échanger +...) en]* $_{Vsup}$ *(augmentation + baisse) de D%*

Les verbes supports transitifs (*Vsupt*) sont exemplifiés en (13) et les verbes supports avec la préposition *en* (*VsupEn*) sont donnés en (14). Les propriétés sont représentées dans les colonnes *N Vsupt Dét V-n De Dnum %* et *VsupEn en Vvn De Dnum %* respectivement.

4.4. *Le composant D%*

D% regroupe les expressions de pourcentage, c'est-à-dire le complément de mesure relative, direct ou indirect.

(15) *(de + E) Dnum (% + pour cent) = (de + E) 15 (% + pour cent)*

4.5. *Le composant DN*

Le composant *DN* est un complément de mesure absolue. Il comporte un déterminant numérique et un nom d'unité monétaire. Il spécifie la valeur des actions échangées à la Bourse.

(16) *Dnum Nunité = 45 euros*

4.6. *Le composant AT*

Les compléments de temps ponctuels qu'on peut trouver dans le corpus sont construits autour des noms d'activités boursières, avec des noms temporels comme *début, milieu*. En voici des exemples :

(17) *(dans les premièr(e)(s) + au début de LE + au cours de LE) (transactions + échanges + cotation)*

Les groupes nominaux prépositionnels comme (17) se comportent comme des adverbes. Mais il n'est pas souhaitable de les étiqueter comme adverbe dans le dictionnaire, parce qu'ils sont composés de locutions prépositives figées donnant le sens de temps et de noms d'activité, qui peuvent varier selon les contextes. Dans ce genre de situation, la construction de la grammaire locale énumérant les noms qui peuvent apparaître dans un corpus particulier est utile.

Des compléments de temps généraux peuvent aussi apparaître dans le corpus. Par exemple :

(18) *dans la matinée du mardi 10 octobre = le mardi 10 dans la matinée*
le mardi 10 = le mardi

Ces différents compléments de temps sont symbolisés par *AT*.

4.7. *Le composant AL*

Les compléments de lieu, symbolisés par *AL,* montrent plusieurs variantes de formes, à cause de la troncation. Dans le corpus, toutes les séquences suivantes sont considérées comme équivalentes :

(19) *à la Bourse de Paris, à Paris, en Bourse, à la Bourse*

4.8. *Le graphe principal*

Pour mener à bien la reconnaissance des phrases dans le corpus, il faut linéaliser les composants décrits plus haut. On appelle graphe principal la grammaire locale qui contient tous les sous-graphes. Les composants *AT, AL, D%* et *DN*, de nature adverbiale, peuvent subir une permutation au sein d'une même phrase, ce qui multiplie le nombre de chemins dans la grammaire locale. Voici le graphe principal appliqué lors du parsing du corpus :

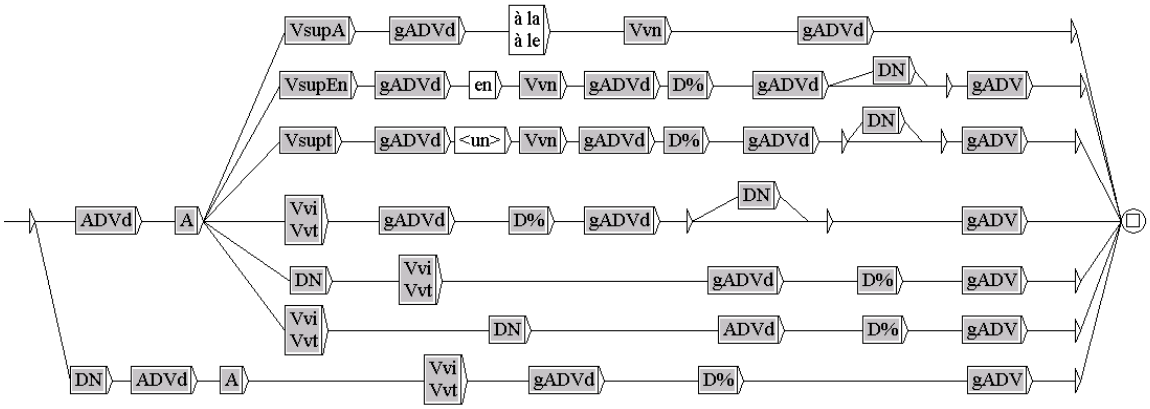


Figure 3. Le graphe principal[5]

Le nombre des sous-graphes dans ce graphe principal est d'à peu près 270[6]. Avec ces grammaires locales, toutes les phrases, y compris celles qui n'ont jamais été rencontrées au cours de la construction de ces grammaires apparaissant dans le corpus étudié qui sont linguistiquement analysables selon le schéma de phrase donné en haut ont été entièrement reconnues sans ambiguïté. Les phrases concernées représentent environ 22 % du corpus total.

Il faut souligner, cependant, le fait que ce résultat a été obtenu en fonction d'un corpus choisi d'avance et que les grammaires locales ont été également conçues spécifiquement pour être appliqué à ce corpus. Il n'est donc pas étonnant que leur application à d'autres corpus, même s'il est de la même nature que celui qu'on a utilisé, ne donne pas, d'emblée, un résultat aussi fin. Or nous pensons que des ajustements mineurs (cf. la différence du style) remettent assez rapidement en marche les grammaires locales conçues ici, puisque les grammaires locales sont modulaires, et qu'il est possible d'atteindre un taux de reconnaissance semblable, du fait que l'information communiquée est de même nature et que nous pensons que la variation de formes linguistiques qui correspond à cette information doit être assez limitée.

Le résultat de l'application de ce graphe au corpus donne une concordance, dont un échantillon est donné ci-dessous (Figure 4) :

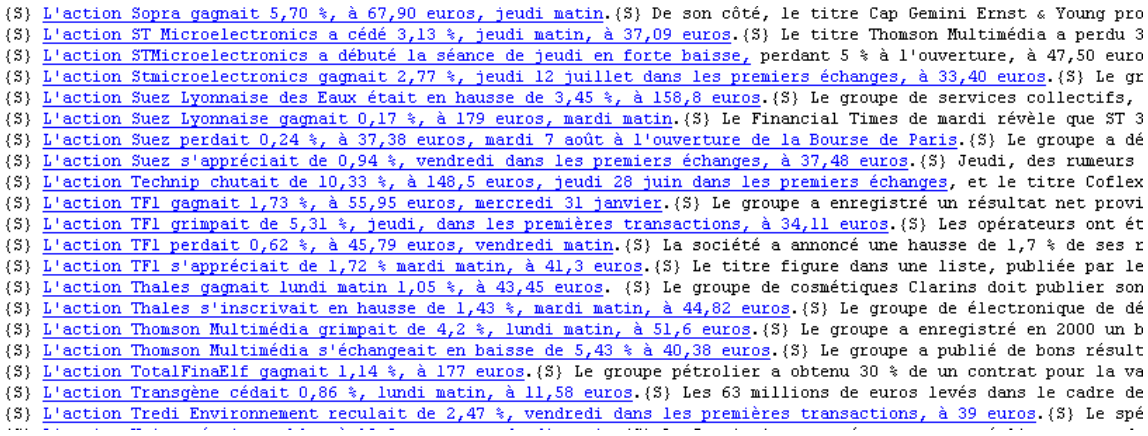


Figure 4. Concordance de la reconnaissance

[5] Une boîte grisée appelle un sous-graphe.
[6] Un sous-graphe peut appeler à son tour un autre sous-graphe.

La partie soulignée correspond à une séquence reconnue par les grammaires locales. La lettre *S* majuscule entre parenthèses signifiant le commencement et la fin d'une phrase, les séquences analysées de la Figure 4 correspondent à des séquences reconnues comme des phrases par *UNITEX*.

5. Conclusion et perspectives

22 % des phrases du corpus sont analysées par le schéma de phrase donné. Le résultat de la reconnaissance automatique n'a pas d'ambiguïté. Cela montre l'utilité d'une analyse linguistique, syntaxico-sémantique et lexicale, préalable à la construction de grammaires locales. Cela est nécessaire pour mener à bien l'analyse automatique de ce type de texte. La technique linguistique de réduction des variantes formelles de phrase à un schéma de phrase par rapport à un discours particulier est une aide pour la création de grammaires locales optimales.

La continuation de ce travail doit tenir compte de plusieurs tâches :

1. pouvoir traiter d'autres types de schémas de phrase, e.g. formaliser le deuxième type de phrase apparaissant dans les rapports boursiers[7] (cf. 3.2.)
2. pouvoir traiter les phrases complexes, e.g. le schéma de phrase étudié ici peut être réutilisé pour décrire par exemple les compléments phrastiques de verbes de parole : *X a annoncé une baisse de 2,1 % de son chiffre d'affaires, à 34 milliards d'euros :* la partie soulignée dans cette phrase peut être formalisée avec les composants que nous avons fixés, complétés par une nouvelle unité, *chiffre d'affaires*.

Cette méthode peut être appliquée à des discours d'autres natures : bulletin météo, résultats sportifs, comptes rendus techniques, etc. On peut cependant prévoir que l'application subirait des contraintes pareilles à celles rencontrées ici, c'est-à-dire que certaines phrases sont faciles à décrire dans des schémas et d'autres sont plus problématiques. Pour le moment, l'objet de description et les résultats sont limités mais nous pensons que l'accumulation de descriptions linguistiques et de grammaires locales, tournées vers des corpus concrets, diminuera ces contraintes.

Références

Gross M. (1975). *Méthodes en syntaxe.* Hermann.

Gross M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Langages*, vol. (63) : 7-52.

Gross M. (1996). Construction de grammaires locales et automates finis. *Working Papers*, vol. (5). Centro linguistico, Universita' commerciale ''L. Bocconi''.

Gross M. (1997). The Construction of Local Grammars. In Roche E. et Schabes Y. (Eds), *Finite State Language Processing.* The MIT Press.

Harris Z. (1952). Discourse Analysis. *Language*, vol. (28).

Harris Z. (1963). *Discourse Analysis Reprints.* Mouton & Co.

Hirschman L. et Sager N. (1982). Automatique Information Formatting of a Medical Sublanguage. In Chapitre 2 de Kittredge R. et Lehrberger J. (Eds), *Sublanguage : Studies of Language in Restricted Semantic Domains*. Walter de Gruyter.

[7] De ces phrases disparates sommairement catégorisées dans ce deuxième type de phrase semble apparaître une échelle de degré de « décrivabilité » de phrases, mesuré en fonction de la nature de l'information. Si la phrase rapporte des événements boursiers, financiers ou commerciaux, elle est plus facile à décrire que la phrase qui rapporte un phénomène naturel, par exemple.

Kittredge R. (1982). Variation and Homogeneity of Sublanguages. In Chapitre 4 de Kittredge R. et Lehrberger J. (Eds), *Sublanguage : Studies of Language in Restricted Semantic Domains*. Walter de Gruyter.

Leclère Chr. (2002). Organization of the lexicon-grammar of French verbs. *Lingvisticae Investigationes*, vol. (25 : 2).

Paumier S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse, Université de Marne-la-Vallée, IGM.

Annexe : Extrait de la table du lexique-grammaire spécialisé[8]

A	**Prédicats de variation**		**Vvi**		**Vvt**	**VsupEn en Vvn De Dnum %**						**Vsup à Dét V-n**						**Vsupt Dét V-n De Dnum %**				
N0 = ACTION	Vv ou Vsup Adj	Vvn	NVDeDnum%	NVDnum%	NVDnum%DePoss0Valeur	être	s'échanger	s'inscrire	pointer	se négocier	Vtps	être réservé	être orienté	s'orienter	reprendre	repartir	pointer	faire un V-n	enregistrer UN V-n	connaître UN V-n	afficher UN V-n	accuser UN V-n
+	apprécier(s')	(appréciation)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	augmenter	augmentation	+	–	–	+	?	?	?	?	+	?	?	?	?	?	?	–	?	?	?	?
+	baisser	baisse	+	–	–	+	+	+	+	+	+	+	?	?	?	?	?	–	+	?	?	?
+	bondir	bond	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	?	–	?	?
+	chuter	chute	+	–	–	?	?	?	?	?	–	–	–	–	–	–	–	?	?	?	?	?
+	décrocher	(décrochage)	+	–	–	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
+	dégringoler	(dégringolade)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	–	–	–
+	déprécier(se)	(dépréciation)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	effondrer(s')	(effondrement)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	effriter(s')	(effritement)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	envoler(s')	(envolée)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	glisser	(glissement)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	–	–	–

[8] Dans cette table du lexique-grammaire spécialisé, le signe ''+'' signifie que l'expression en question est trouvée dans le corpus. Le signe ''–'' montre le cas contraire. Le signe d'interrogation ''?'' signifie que l'expression en question n'est pas agrammaticale, mais qu'elle n'a pas apparu dans le corpus.

A	Prédicats de variation		Vvi		Vvt	VsupEn en Vvn De Dnum %						Vsup à Dét V-n						Vsupt Dét V-n De Dnum %				
N0 = ACTION	Vv ou Vsup Adj	Vvn	NVDeDnum%	NVDnum%	NVDnum%DePoss0Valeur	être	s'échanger	s'inscrire	pointer	se négocier	Vtps	être réservé	être orienté	s'orienter	reprendre	repartir	pointer	faire un V-n	enregistrer UN V-n	connaître UN V-n	afficher UN V-n	accuser UN V-n
+	grimper	(grimpée)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	progresser	progression	+	–	–	+	?	?	?	?	+	?	?	?	?	?	?	–	+	?	?	?
+	rebondir	(rebond)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?	?
+	reculer	recul	+	–	–	+	?	?	?	?	?	–	–	–	–	–	–	–	+	?	+	+
+	replier(se)	repli	+	–	–	+	?	?	?	+	+	?	?	?	?	?	?	?	?	?	+	?
+	reprendre(se)	(reprise)	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?	?
+	stable(être + rester)		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+		hausse	–	–	–	+	+	+	+	+	+	+	+	+	+	+	+	?	+	+	+	?
+	abandonner	(abandon)	–	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	adjuger(s')		–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	céder	(cession)	–	+	+	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?
+	gagner	gain	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	+	?
+	perdre	(perte)	–	+	+	–	–	–	–	–	–	–	–	–	–	–	–	?	?	?	?	?
+	prendre		–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
+	regagner	(regain)	–	+	–	–	–	–	–	–	–	–	–	–	–	–-	–	–	?	?	?	?

Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé

Berthille Pallaud[1], Sandrine Henry[2]

[1] UMR 6057 – Université de Provence – Aix-en-Provence – 13621 cedex 1 France
[2] DELIC – Université de Provence – Aix-en-Provence – 13621 cedex 1 France
berthille.pallaud@lpl.univ-aix.fr, sandrine_henry@hotmail.com

Abstract

Word fragments are, like repeats, very frequent events in spoken language. Speakers and listeners alike are generally unaware of these phenomena as if they were not part of the communication process. Previous studies (Pallaud, 2002; Henry, 2002; Henry and Pallaud, 2003) have shown that word fragments and repeats encompass different categories of phenomena, and this enables us to define them as an heterogeneous group ruled by different types of constraints and mechanisms. These two types of disfluences may be considered less as repairs than hesitation phenomena. Fragments of word differ from repetitions: they occur later in the syntagm and in the clause.

Résumé

Les amorces ou fragments de mots tout comme les répétitions sont des événements langagiers d'une grande fréquence à l'oral. Ces deux phénomènes sont toutefois le plus souvent ignorés du locuteur qui parle et de son récepteur comme si ces phénomènes étaient non-communicatifs. Des travaux antérieurs (Pallaud, 2002 ; Henry, 2002 ; Henry et Pallaud, 2003) ont montré que les amorces de mots comme les répétitions regroupent des phénomènes de catégories différentes ce qui permet de les caractériser comme un ensemble hétérogène obéissant à plusieurs types de contraintes ou de mécanismes.
L'analyse de ces deux sortes d'achoppement montre qu'ils sont moins le moment d'une réparation d'erreurs que celui d'une hésitation dans l'élaboration de l'énoncé. La comparaison des amorces de mots et des répétitions fait apparaître une nette différence : les répétitions se produisent plus tôt dans le syntagme que ne le font les amorces de mots.

Mots-clés : amorces de mots, répétitions, disfluences, français oral et spontané

Keywords : word fragments, repeats, disfluencies, oral and spontaneous French

1. Introduction

Notre étude s'intéresse à ces phénomènes caractéristiques de l'oral et souvent qualifiés de « scories » ou même « ratés » : les amorces de mots et les répétitions. Dans notre étude, le terme d'amorce (ce que Schegloff *et al.* (1977) nomment des *auto-corrections*) est conforme à l'emploi qui en est fait dans les conventions de transcription du *Groupe Aixois de Recherche en Syntaxe* (GARS) : il est réservé aux ratés dans la production de morphèmes. Les fragments de syntagme sont, quant à eux, plus souvent identifiés sous le terme de *bribes* (Cappeau, 1998). Il reste que ces trébuchements ne sont pas toujours distingués. Ils sont parfois analysés comme une seule catégorie de fragments (Jeanjean (1984) et Martinie (1999) pour ce qui est des études sur le français parlé ; Eklund (2003) en ce qui concerne d'autres langues). Les conventions de transcription du GARS permettent un repérage automatique de ce type d'achoppement :

FEMMES 51,15 pour moi c'est pas *malho-* c'est pas *malhonnête* c'était euh

Dans la très grande majorité des cas, ces phénomènes étant identifiés comme des « réductions » de mots qui ne sont qu'amorcés, leur transcription est orthographique et atteste ainsi de cette interprétation.

Même si les études de ces phénomènes particuliers à l'oral ne les catégorisent pas toujours de façon semblable, elles constatent toutes que ces achoppements ne sont pas rares et ne se produisent pas de façon aléatoire (Duez, 2001b ; Pallaud, 2002a). En moyenne (sur un corpus de 46 549 mots[1]), on peut dire qu'on trouve environ 4 amorces tous les 1000 mots soit pour un débit moyen de 200 mots par minute une amorce toutes les 75 secondes. Avec un corpus de un million de mots (*conf. infra*), cette fréquence est encore plus grande puisqu'on retrouve une amorce toutes les 50 sec (Henry et Pallaud, 2003d). Ces phénomènes viennent loin derrière les pauses courtes silencieuses ou remplies qui sont six fois plus nombreuses (Pallaud, 2002a).

Une analyse comparative de ces douze corpus (un locuteur par corpus) a été faite sur *le nombre moyen de mots entre deux amorces* afin d'évaluer le nombre d'événements (ici les amorces) tous les 1000 mots. Elle montre qu'il existe une grande hétérogénéité entre les locuteurs puisque les valeurs peuvent varier d'un facteur 10. Cette variabilité observée dans la production des amorces suggère que ce paramètre reflète fortement une composante individuelle. Le phénomène d'amorce semble être une marque d'énonciation qui rend compte de caractéristiques subjectives : « Les hésitations peuvent fonctionner comme des marqueurs de personnalité » (Duez, 2001 : 129).

Notre étude sur les amorces répétées (Henry et Pallaud, 2003c) a confirmé, quant à elle, que la distinction entre les amorces et les répétitions de mots s'imposait. Elle a suggéré, de plus, que les répétitions de mots se produisent plus tôt que les interruptions de mots dans le cours du syntagme. Reste à savoir si cette préséance se retrouve également au niveau de la proposition (avant plutôt qu'après le verbe, par exemple). Nous avons constaté, par ailleurs que le phénomène d'amorce a été souvent considéré comme une auto-correction, un phénomène de réparation ou d'hésitation. Notre analyse de ces achoppements se propose de préciser en quoi ils sont une manifestation d'hésitation plutôt qu'une erreur d'énonciation autrement dit une pause dans l'énoncé plutôt qu'un trébuchement auquel le locuteur remédiera.

2. Corpus et méthodes

À partir des années 1970, sous l'impulsion et la direction de Jean Stéfanini puis de Claire Blanche-Benveniste, de nombreuses études sur le français parlé ont été entreprises dans le cadre du *Groupe Aixois de Recherche en Syntaxe*. Elles ont nécessité l'établissement de corpus oraux dont l'objectif était de recueillir des données représentatives d'un français parlé qu'il serait possible de qualifier d'un « usage général et courant ». Un premier ensemble a été constitué connu sous le nom de *Corpaix*

Depuis les années 1990, toutes les transcriptions ont été saisies sur ordinateur et mises à un format standard. Sont distinguées la partie « ancienne » du corpus (non-informatisée) et la partie « moderne » qui comprend systématiquement trois supports : enregistrement sonore (sur minidisk, puis transfert sur ordinateur), transcription sur papier et saisie informatisée du texte transcrit. La partie « moderne » comporte, au début de l'année 2000, environ 1 million

[1] Cet ensemble est composé de 12 corpus recueillis par Pallaud dans le cadre du *Corpus de Référence de Français Parlé* (Pallaud, 2001). Par ailleurs, l'ensemble de la base de données sur les amorces est tirée du recueil *Corpaix* (une série de corpus d'entretiens référés par un nom) ; cet ensemble a été interrogé à l'aide d'un logiciel concordancier (*Contextes*) élaboré par Véronis (http://up.univ-mrs.fr/~veronis/).

de mots. Dans cet ensemble, ont été retenus des locuteurs ayant fait leur scolarité primaire en France. Le GARS ne privilégie pas l'étude des productions dites « spontanées » (conversations ou entretiens libres), un grand nombre de situations « contraintes » ont été retenues (demandes d'explications ou de renseignements, exposés, descriptions et récits etc.), avec des lieux très variés : domicile d'une personne, lieu de travail, magasin, bureau, organisme public ou radio et télévision. Aucun enregistrement n'a été fait en laboratoire.

L'ensemble du corpus, enregistrements et transcriptions, a été fait, pour la plupart, grâce au travail d'étudiants. Deux exigences, souvent peu compatibles, ont prévalu : celle de la transcription, qui doit être fidèle (noter tout ce qui est dit, y compris les répétitions, les bribes, etc.) et la nécessité de rester lisible, au moins pour pouvoir être lue facilement par une communauté de chercheurs et pour ne pas rester confinée au groupe qui l'a réalisée. Comme le GARS étudie essentiellement des phénomènes de syntaxe qui se saisissent sur de longs passages, où la lisibilité est primordiale, le choix est celui d'une transcription « pauvre », sans surcharge pour noter les mimiques, les rires, les toux, les phénomènes d'interaction gestuelle, ni même les intonations.

Les transcriptions sont faites en orthographe standard, sans aucun « trucage ». Le modèle orthographique est celui des dictionnaires, y compris pour les interjections (*euh, ben*, etc.). Des conventions particulières interviennent pour les éléments qui ne peuvent pas recevoir une orthographe française normée : mots étrangers, néologismes, etc. Les transcriptions ne comportent aucune ponctuation. En 1995, les nouvelles technologies, employées au sein de l'ESA-CNRS 6060, ont rendu possible le regroupement des transcriptions (le corpus *Corpaix*) et le développement d'un logiciel d'exploitation (Xcor un concordancier mis au point par JP Adam à partir de 1997).

Le « Corpus de référence de français parlé » répond à une requête de la Délégation à La Langue Française, Ministère de la Culture, qui l'a totalement financé. La réalisation de ce projet avait été confiée en 1998 à l'URA CNRS 6060, à partir de 2000, le projet a été pris en charge par la Jeune équipe DELIC (*DEscription Linguistique Informatisée sur Corpus*), dirigée par Jean Véronis. L'objectif de ce corpus était de mettre à la disposition de la communauté de linguistes des témoignages de la langue française parlée aujourd'hui, dans les principales villes de l'hexagone. Des enregistrements (132) de locuteurs choisis selon des critères précis (âge, niveau d'éducation, etc.), dans une quarantaine de villes de France, représentant des types de parole variés (parole publique, professionnelle ou privée) sont transcrits selon les normes du GARS. La totalité[2] est estimée à environ 50 heures de parole, soit environ 400 000 mots. La transcription et la segmentation ont été réalisées majoritairement à l'aide du logiciel *Transcriber 1.4.2.*[3], programme développé par Claude Barras spécialement pour la transcription et l'alignement de corpus oraux.

Une fois l'intégralité du corpus obtenue, la suppression des balises XML et la concaténation de l'intégralité du corpus ont été effectuées grâce à des scripts en Perl afin de pouvoir effectuer des requêtes avec des concordanciers classiques. Le corpus a donné lieu au développement de plusieurs outils, notamment d'un concordancier (*Contextes*, développé par Jean

[2] Pour le français parlé, il offre une base de comparaison avec les corpus de français parlé hors hexagone, le corpus Valibel en Belgique (dix fois plus important) et le corpus d'Ottawa-Hull au Canada ; il représente l'amorce de la réalisation d'un corpus comparable à celui du British National Corpus, pour l'anglais parlé.

[3] Logiciel téléchargeable gratuitement sur le site : http://www.etca.fr/CTA/gip/Projets/Transcriber/

Véronis) qui permet d'effectuer des requêtes sous forme de chaînes de caractères et d'obtenir la concordance sous forme textuelle.

La présente étude prend appui sur des données extraites de *Corpaix* auquel s'adjoignent des transcriptions inédites pour former un corpus de travail de 1 000 382 mots (283 situations, 794 locuteurs). Afin d'assurer une fiabilité dans l'extraction des répétitions, celles-ci sont détectées à l'aide d'un programme informatique : écriture d'un script en langage *Perl*, puis application sous *Linux*. L'extraction des amorces s'est faite à l'aide du concordancier Xcor sur les mille premières amorces. L'ensemble des résultats sont ouverts dans un tableur *Excel* pour en faciliter le traitement.

3. Analyse des fonctions de l'amorce et de la répétition de mots

Pour tester notre hypothèse selon laquelle les phénomènes d'amorces ou de répétitions de mots sont moins le moment d'une erreur que la marque d'une élaboration de l'énoncé, nous avons retenu plusieurs paramètres : le poids respectif des trois types d'amorces, les effets de la répétition et de l'insertion d'éléments sur les amorces.

3.1. *Les types d'amorces*

Les amorces ne sont pas toutes semblables. Nous avons identifié (par annotation manuelle) trois sortes d'amorces dont la distinction repose sur l'effet de l'interruption du mot au lieu syntaxique qu'il occupe. Y a-t-il réduction définitive sur cette place syntaxique, auquel cas l'amorce n'est pas complétée ni modifiée mais simplement laissée inachevée, et ce qui suit l'amorce occupe une autre place syntaxique ? Si ce qui suit l'amorce ne change pas de place syntaxique, deux cas sont possibles. Y a-t-il piétinement sur la place syntaxique et une simple reprise puis poursuite du morphème inachevé ou plutôt une modification lexicale ? La réponse à ces questions permet de distinguer trois catégories d'amorces :

– **amorce laissée inachevée**
  FEMMES 10,9 je pense que par rapport à M. C. on a souvent *tendan-* à à analyser un peu les choses euh à dire

– **amorce complétée** (dont celles qui s'accompagnent de <u>modifications antérieures</u>)
  METALL 21,16 donc le cours *cons- consiste* en gros euh à à un exposé théorique
  BUSABCD 20,3 elles suivent des cours d'aide-soignante <u>*en*</u> *d-* euh <u>*au bout de*</u> *dix* mois elles sortent diplômées

– **amorce modifiée**
  ARBORIGN 13,4 pour les bras j'ai déjà *dét- retravaillé* les coudes et et allongé les bras

Les amorces complétées dans un corpus de plus d'un million de mots sont largement majoritaires (60 %). Viennent ensuite celles qui sont inachevées (22 %) et modifiées (18 %). Ce fait nous incite à conclure que, dans les trois quart des cas (82 %), l'amorce de mot est un achoppement marquant l'hésitation plus que l'erreur de langage. L'amorce complétée ou inachevée prélude le plus souvent à une poursuite de l'énoncé sans qu'une modification intervienne. L'étude de Cappeau (1998) sur les phénomènes de bribes à l'oral (c'est-à-dire des amorces répétées de syntagmes et non pas de mots) en relation avec le genre a montré que les bribes qui s'accompagnent d'un changement de genre sont 5 à 8 fois moins nombreuses que les précédentes. Son résultat confirme donc les nôtres : les bribes complétées sont beaucoup plus fréquentes que les bribes modifiées.

On constate par ailleurs (Figure1) que cette répartition est très différente selon que ces achoppements sont recueillis *à la volée* ou recherchés dans des corpus enregistrés (ici *Corpaix* ou le corpus *Bertuzzi97*[4]) :

Amorces complétées : deux exemples de ce type seulement (soit 5 % des amorces) se trouvent dans les amorces relevées *à la volée*, alors qu'ils abondent dans les corpus enregistrés (*Corpaix* et *Bertuzzi97*) où ce type d'amorce représente respectivement 60 % et 30 % des amorces émises.

Amorces laissées inachevées : dans le corpus *à la volée*, on ne relève aucune amorce de ce type, alors que dans les corpus enregistrés, ces amorces inachevées constituent près de 20 %.

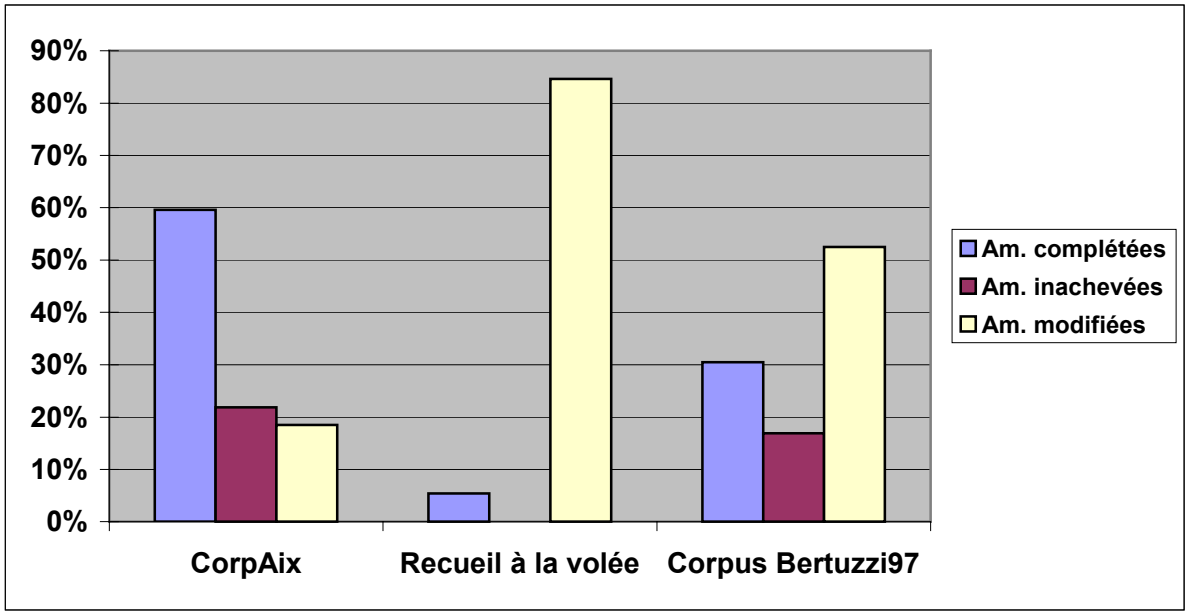


Figure 1. Répartition des trois types d'amorces en fonction du type de recueil

Amorces modifiées : ces phénomènes de modification sur l'élément touché par l'amorce représentent la quasi-totalité des amorces relevées *à la volée* alors que la proportion est de 18 % seulement dans le grand corpus *Corpaix* et 52,5 % dans le corpus *Bertuzzi97*.

Il n'y a donc pas que la fréquence des amorces qui peut varier d'un locuteur à l'autre. La répartition des trois types d'amorces peut également différer de façon non négligeable. La comparaison des résultats obtenus dans un gros corpus (*Corpaix*) avec les amorces relevées pour un locuteur (le corpus *Bertuzzi97*) confirme qu'il y a probablement une grande variabilité d'un corpus à l'autre c'est-à-dire d'un locuteur à l'autre.

Ces résultats montrent que, dans un recueil *à la volée,* non seulement les amorces de mots sont sous-représentées dans les « erreurs » de langage mais seules les amorces suivies d'une modification sont véritablement détectées. Comme, en fait, les amorces modifiées produites par les locuteurs ne sont pas les plus nombreuses (18,5 % dans notre grand corpus), on peut penser que les amorces complétées et les amorces inachevées ne sont pas perçues ou ne sont pas considérées comme des « erreurs » de langage. Sur le versant de la perception, ces trébuchements ne seraient pas disfluents, du moins à ce stade de la discussion.

[4] Le corpus *Bertuzzi97* fait partie de l'ensemble *Corpaix* ; il comprend 25 460 mots et 100 amorces de mots.

3.2. *Les effets en contexte de l'achoppement*

D'autres résultats viennent soutenir cette hypothèse. Il s'agit des effets en contexte de l'achoppement que constitue le phénomène d'amorce. Trois sortes de phénomènes seront pris en compte : les modifications antérieures à l'amorce de mot, les insertions d'éléments à l'occasion de phénomènes d'amorces et le cas des amorces répétées.

3.2.1. *Modifications antérieures à l'amorce de mot.*

Parmi les cas d'amorces complétées, toutes ne s'accompagnent pas d'une simple reprise. Nous avons en effet trouvé un certain nombre de modifications qui portent sur des éléments antérieurs dans l'énoncé. Le locuteur arrête le cours de son énoncé au milieu d'un mot, reprend une partie de son énoncé et en corrige un élément :

> HISTOIRE 3,14 comment on reconnaît un Belge **dans** un marchand de *chau-* **chez** un marchand de *chaussures*

L'amorce est bien complétée mais une modification intervient sur un élément qui a déjà été prononcé. L'amorce n'est donc pas le lieu de la réparation même si elle « annonce » dans ces cas une modification. Ces achoppements ne sont pas des amorces modifiées puisqu'en fait elles seront complétées après qu'une partie de l'énoncé antérieur ait été remplacé. Parfois, la modification antérieure consiste à ôter une élément présent dans l'énoncé précédent. Cette catégorie est rare ; 9 cas seulement sur les 1 000 amorces étudiées :

> PUGET-VI 28,9 mais souvent c'est que + tous les **petits** *en-* + tous les *enfants* c'est le désir qu'ils ont de jouer

Ces cas de modification antérieure sont réservés aux amorces complétées et ne sont pas fréquents ; ils sont au nombre de 44 dans le corpus *Corpaix* et représentent 4,6 % du total des amorces (44/948) et 7,7 % (44/565) des amorces complétées. Rien ne s'opposerait à ce que ce type de phénomène soit observé également dans le cas des amorces modifiées mais, en fait, nous n'en avons observé aucun. Par ailleurs, ces résultats incitaient à étudier jusqu'où le locuteur peut remonter dans son énoncé et produire une réparation. Il ne s'agit pas de déterminer un empan de la réparation antérieure mais de voir si cette modification « rétroactive » dépasse les limites du syntagme engagé. On s'aperçoit que sur les 44 amorces de cette catégorie, seules quatre modifications concernent un syntagme antérieur différent de celui où se trouve l'amorce « annonciatrice ». Encore constate-t-on qu'éléments amorcés et éléments modifiés sont tous (sauf un) dans la rection du même verbe :

> BEAUMETT 2,15 quand les **gens** vont à *l'arm-* les **hommes** vont à *l'armée*

3.2.2. *Les insertions d'éléments à l'occasion de phénomènes d'amorce*

Dans d'autres cas encore, l'amorce n'est pas le lieu-cible de la réparation. Elle est suivie d'une reprise d'une partie de l'énoncé dans lequel se trouvent insérés un ou plusieurs éléments :

> BENOT 3,13 tu vois + après nous *on li- on* **les** *livre* comme ça

Cette sous-catégorie n'est pas négligeable puisqu'elle accompagne un tiers des amorces (29,5 %), toutes catégories confondues. De plus, ces insertions ne se produisent pas avec la même fréquence selon les types d'amorces : 71,2 % sur les amorces complétées (210/295), le cinquième sur les inachevées (19 % ; 56/295) et 9,8 % sur les modifiées (29/295). La comparaison des pourcentages d'amorces complétées, inachevées et modifiées dans le grand corpus *Corpaix* avec la répartition de ces insertions selon ces types d'achoppements montre une dif-

férence significative ($Khi^2 = 13,58$; d.d.l. = 2 ; $p<.01$). Les insertions sont plus fréquentes lors d'amorces complétées et moins fréquentes lors d'amorces modifiées. Ces ajouts à l'occasion d'interruptions de mots est particulièrement la marque d'une activité autonymique puisque les amorces concernées ne sont pas le lieu d'une modification mais annoncent une ré-élaboration sur ce qui a été énoncé antérieurement.

Nous avons distingué trois sortes d'insertions pouvant avoir lieu à l'occasion de ces achoppements : les insertions « énonciatives », les insertions qui « modulent » le texte et les incises parenthétiques. Les insertions « énonciatives » sont les plus nombreuses (62 % ; 183/295) :

> DESSININ 15,2 alors là on voit le l'aspect euh de la *com-* **enfin** l'organe de *commande*

viennent ensuite les insertions qui modulent (28,1 %) :

> PORTUGA4 10,3 c'est vraiment de voir les choses *très différen-* **d'une façon** *très différente*

et les incises parenthétiques (9,8 %) :

> BARBARA 1,15 tu allais à la boucherie *si tu av-* **certaines boucheries** *si tu avais* de l'argent

L'analyse de la répartition de ces trois types d'insertions selon les types d'amorces montre des variations très nettes. Les insertions « énonciatives » constituent presque la moitié des insertions lors d'amorces complétées (47,6 % ; 100/210).

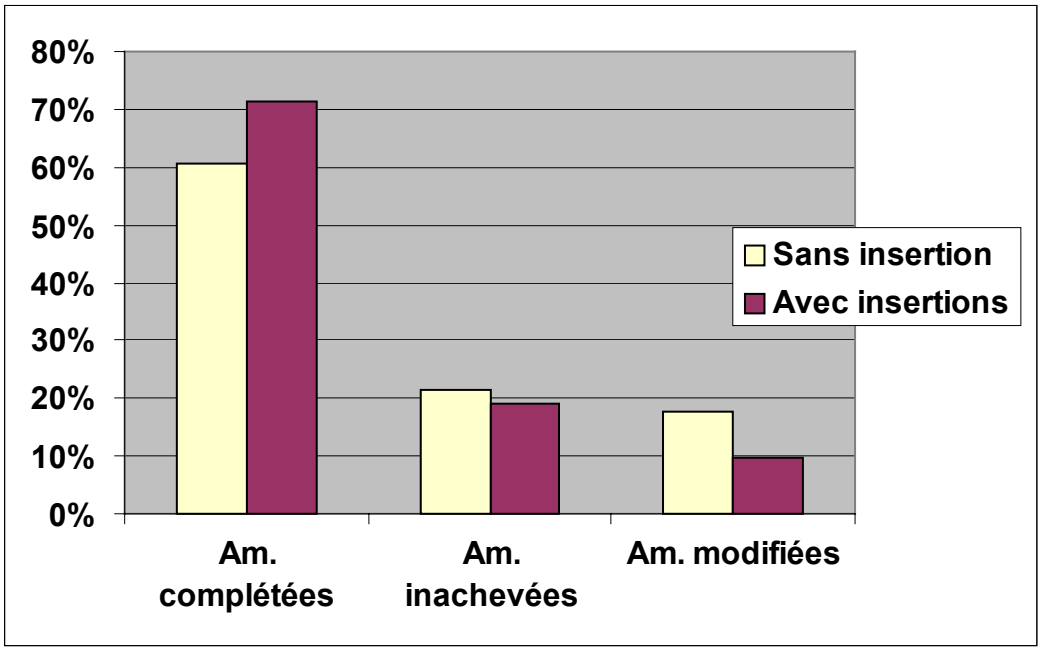


Figure 2. Répartitions des insertions d'éléments en fonction des types d'amorces.

Elle sont également les seuls ajouts pour les amorces inachevées ainsi que la quasi-totalité pour les amorces modifiées. La catégorie d'insertions non « énonciatives » (simples ajouts et incises) semble donc réservée aux amorces complétées.

3.2.3. *Le cas des amorces répétées*

Dans une étude récente (Henry et Pallaud, 2003d), nous avons montré que les amorces répétées constituaient sur certains points une catégorie particulière à la fois d'amorces et de répétitions :

NAVALE 14,11 et que on le *re- re-* revendique encore une fois sur le cahier mais euh

Cette approche contrastive (au regard des amorces et des répétitions) permet de faire des hypothèses notamment sur une fonction probable de la répétition. En effet, on retrouve dans les amorces répétées les trois types d'amorces (complétées, inachevées et modifiées). La répartition de ces amorces est affectée significativement par la répétition ($Khi^2 = 29,85$; d.d.l. = 2 ; $p<.001$) : en effet, les amorces répétées complétées sont de loin les plus nombreuses (82 %) et ce pourcentage est significativement plus élevé que dans le cas des amorces en général. Par ailleurs, les amorces inachevées sont deux fois moins nombreuses (13,3 %) et les amorces modifiées sont quatre fois moins fréquentes (4,7 %).

Ces résultats attestent de la relation de dépendance qui existe entre la catégorie d'amorce, quelle qu'elle soit, et le phénomène de répétition. Cela suggère une fonction de la répétition dans cet achoppement que constitue le phénomène d'amorce : elle soutient la recherche lexicale manifestée par l'amorce et contribue à l'aboutissement de cette recherche puisque, finalement, le mot interrompu est, alors, plus souvent complété et beaucoup moins souvent modifié. Cette dépendance entre répétitions et amorces suggère également que ces répétitions d'amorces pourraient être équivalentes à des pauses remplies ou silencieuses.

4. Place de l'amorce et de la répétition dans le syntagme

Blanche-Benveniste (2003) soutient que la répétition s'attache plus souvent au début du syntagme. Sandrine Henry (2002) confirme cette hypothèse. Qu'en est-il pour l'amorce ? Intervient-elle plus tardivement ou non que la répétition dans le syntagme ? Par ailleurs, les types d'amorces touchent-ils de façon semblable la zone de texte avant ou après le verbe ?

4.1. *Amorces, répétitions et zones de texte.*

Même dans les cas où on ne peut interpréter avec certitude toutes les amorces, leur place syntaxique est connue plus sûrement. Comme le soulignait Jeanjean (1984 : 171) qui conteste le terme de *ratés* pour ces phénomènes : « *Pour assigner une valeur syntaxique à ce genre d'amorces, on n'a pas besoin de reconstituer un quelconque lexème* ». Nous nous sommes intéressées à la façon dont amorces et répétitions se situent dans le texte de l'énoncé : sur le verbe ou dans les zones pré- ou post-verbales (Figure 3).

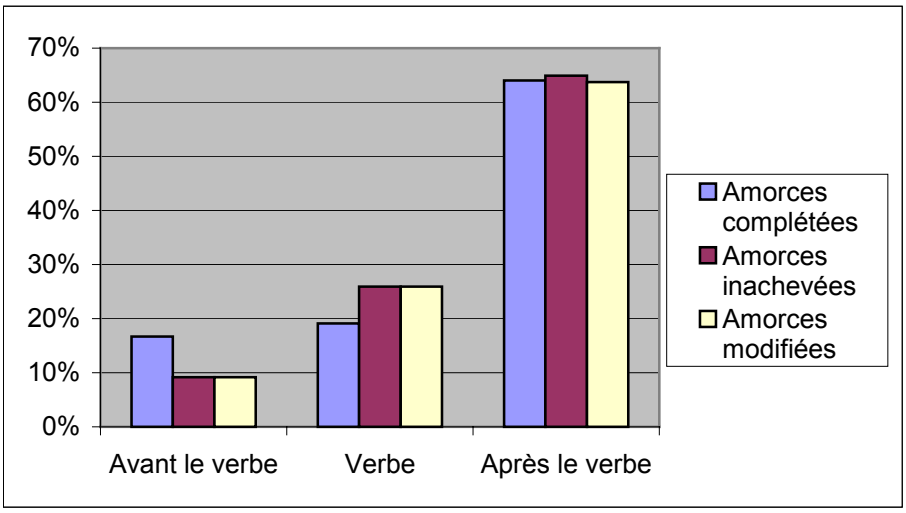


Figure 3. Types d'amorces et zones de texte (verbe, pré- et post-verbale)

La comparaison des localisations occupées par les amorces de mots (Pallaud, 2002a) et les répétitions[5] a suggèré une dépendance[6] entre le type d'achoppement et la zone de texte où les hésitations se situent. Si l'on admet que les zones-textes pré- et post-verbales sont de taille semblable, ces achoppements apparaissent tous deux plus fréquemment dans la zone post-verbale (64 % pour les amorces et 54 % pour les répétitions). En revanche, dans la zone verbale, les amorces sont plus fréquentes que les répétitions (21,9 % *vs* 11,3 %) ; dans la zone pré-verbale, les fréquences sont inverses : 13,9 % pour les amorces et 33,8 % pour les répétitions. Ce résultat suggère que, dans la zone pré-verbale, les répétitions sont plus fréquentes que les amorces. Comparativement aux amorces, les répétitions qui se produisent plus tôt dans le syntagme, sont observées, également, plus tôt dans l'énoncé.

Une étude a été entreprise pour évaluer si les trois types d'amorces touchaient de façon similaire les zones verbales, pré-verbales et post-verbales (Figure 3). Qu'ils soient complétés, inachevés ou modifiés, lorsque les éléments amorcés sont à la droite du verbe, ils représentent un pourcentage semblable : 63 %. Sur les zones verbale et pré-verbale, il existe une différence significative ($Khi^2 = 14, 25$; d.d.l. = 2 ; $p<.001$.) entre les amorces complétées et celles qui ne le sont pas. Une amorce sur quatre touche le verbe et sera modifiée ou restera inachevée alors que l'amorce complétée n'affecte un verbe que dans un cas sur cinq. En revanche, celle-ci concerne la zone avant le verbe dans deux fois plus de cas que l'amorces inachevée ou modifiée. Ainsi, parmi les amorces, celles qui sont complétées sont plus fréquemment situées sur la zone pré-verbale, et moins nombreuses en zone verbale. Lorsque les amorces sont en zone post-verbale, les proportions de chacune des catégories sont semblables.

4.2. *Les catégories grammaticales touchées par les amorces*

Notre étude (Pallaud et Henry, 2003c) a montré que les unités lexicales ne sont pas affectées de la même manière par les amorces et les répétitions. Il apparaît utile de distinguer les mots qui ont une charge lexicale pleine, les *mots pleins*, de ceux qui participent de la structuration de la langue : les *mots-outils*. Les résultats obtenus pour les répétitions (Corpus *Corpaix* : 15 985 répétitions) et les amorces (999 amorces analysées) ont révélé une différence très nette. Les répétitions touchent plus souvent (91,3 %) les mots-outils que les amorces qui affectent plus fréquemment les mots pleins (67,9 %).

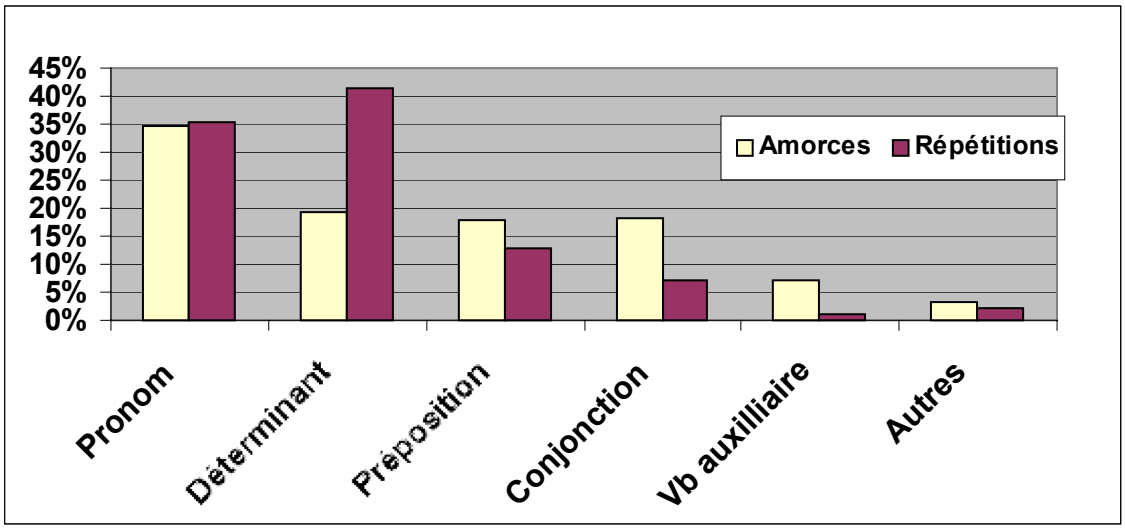


Figure 4. Types de mots-outils dans les amorces de mots et les répétitions.

[5] Corpus *2000-EDF* : 15 751 mots (393 hésitations réparties : 8 sur place « associée »).

[6] $Khi^2 = 74,86$; d.d.l. = 2 ; $p<.001$; il reste pour être affirmatif à tenir compte de la taille des zones-textes avant et après le verbe, ce qui n'a pas été encore fait.

Sur la totalité des amorces, celles qui concernent des « petits mots » ou des pronoms sont de loin les moins nombreuses (22,1 %). Que cela soit dû à un lien entre la taille du mot et le phénomène d'amorce resterait à vérifier. Dans les syntagmes nominal et verbal, les mots pleins ne sont le plus souvent qu'en deuxième position sur l'axe syntagmatique. Les pronoms ou les déterminants les précèdent dans la très grande majorité des cas. Les répétitions affecteraient donc, plus tôt que les amorces, le syntagme nominal ou verbal.

Parmi les mots-outils, deux catégories grammaticales sont majoritairement concernées lors de la répétition : il s'agit des déterminants et des pronoms (respectivement 41,5 % et 35,5 %). Viennent ensuite les prépositions et locutions prépositionnelles (13 %), ainsi que les conjonctions et locutions conjonctives (7 %). En ce qui concerne les amorces, la répartition est différente ($Khi^2 = 25,46$; d.d.l. = 2 ; $p<.001$.) : la catégorie des pronoms est la plus touchée avec 34,8 %. Les amorces affectent deux fois moins les déterminants mais deux fois plus les conjonctions et locutions conjonctives. Henry (2002a) a trouvé sur les formes *le, la les* et *leur* que les déterminants font nettement plus souvent l'objet d'une répétition que les pronoms clitiques compléments. Ces résultats et les précédents encouragent à conclure que la répétition s'attache plus souvent au début du syntagme alors que l'amorce qui touche plus fréquemment le lexique affecte le syntagme en un point plus avancé.

Conclusion

Les amorces de mots comme les répétitions ont en commun une suspension de l'énoncé sur une place syntaxique. Ces deux phénomènes, tous deux fréquents à l'oral, sont loin d'être des procédés énonciatifs, tant ils échappent, le plus souvent, à la perception même des locuteurs et de leurs « récepteurs » (Pallaud, 2003a ; Roubaud, 2003). Leur fréquence comme leur catégorie peut varier d'un locuteur à l'autre. L'exploration des effets en contexte des phénomènes d'amorces montre que ces trébuchements sur un morphème ne constituent pas « l'erreur » elle-même mais l'indice qu'une modification va intervenir sur l'énoncé déjà prononcé. Dans moins d'un cas sur cinq, le morphème interrompu fera l'objet d'une modification. Tous les autres cas montrent que le morphème est repris pour être complété (60 %) ou abandonné (22 %).

Par ailleurs, dans un tiers des cas, toutes catégories d'amorces confondues, ces achoppements s'accompagnent d'une insertion ou d'une reprise de l'énoncé antérieur qui sera alors modifié. Le type d'amorce très majoritairement touché est l'amorce qui sera finalement complétée (74,9 %). Outre que le cas de l'amorce complétée est le plus fréquent, il semble, au regard des insertions constituer une catégorie particulière : il est le seul (à deux exceptions près) à s'accompagner d'insertions qui modulent ou d'incises parenthétiques. Ces résultats suggèrent que l'usage du terme de *reparandum* (Shriberg, 1999) pour qualifier le morphème interrompu ou le terme qui va être répété est sans doute inapproprié. Dans la plupart des cas, ce n'est pas sur cet élément que va avoir lieu une modification.

Si on examine la localisation des achoppements sur les zones verbales, pré- et post-verbales, on observe des différences entre amorces et répétitions mais également entre catégories d'amorces. Comparativement aux amorces, les répétitions qui se produisent plus tôt dans le syntagme, sont observées, également, plus tôt dans l'énoncé. Parmi les amorces, celles qui sont complétées sont plus fréquemment situées sur la zone pré-verbale, et moins nombreuses en zone verbale. Lorsque les amorces sont en zone post-verbale, les proportions de chacune des catégories sont semblables.

Les répétitions touchent, bien plus souvent que les amorces, les mots-outils tandis que les amorces affectent, plus fréquemment que les répétitions, les mots pleins. L'examen des catégories grammaticales affectées par les répétitions et les amorces nous semble permettre de conclure que la répétition s'attache plus souvent au début du syntagme alors que l'amorce qui touche plus fréquemment le lexique affecte le syntagme en un point plus avancé.

Références

Blanche-Benveniste C. (2003). La naissance des syntagmes dans les hésitations et les répétitions du parler. In Araoui J.L. (Ed.), *Le sens et la mesure. Hommages à Benoît de Cornulier.* Honoré Champion : 40-55.

Cappeau P. (1998). Quelques mots sur quelques bribes liées au genre. In Bilger M., Van den Eynde K. et Gadet F. (Eds), *Analyse linguistique et approches de l'oral. Recueil d'études offert en hommage à Claire Blanche-Benveniste.* Peeters : 301-311.

Duez D. (2001 b). Signification des hésitations dans la production et la perception de la parole spontanée. *Revue Parole*, vol. (17-18-19) : 113-138.

Eklund R. (2003). *Disfluency in Spontaneous Speech Workshop*. In *Proceedings of DiSS'03*, 5–8 September 2003, Göteborg University.

Henry S. (2002). Étude des répétitions en français parlé spontané pour les technologies de la parole. In *Actes de RECITAL 2002* : 467-476.

Henry S. et Pallaud B. (2003c), Amorces de mots et répétitions dans les énoncés oraux. *Recherches Sur le Français Parlé,* vol. (18). A paraître

Henry S. et Pallaud B. (2003d). Word fragments and repeats in spontaneous spoken French. In Eklund R. (Ed.), *Disfluency in Spontaneous Speech Workshop*, in *Proceedings of DiSS'03*, 5–8 September 2003, Göteborg University : 77-80.

Jeanjean C. (1984). Les ratés c'est fa- fabuleux. Étude syntaxique des amorces et des répétitions. *Linx*, vol. (10) : 171-177.

Martinie B. (2000). Remarques sur la syntaxe des énoncés réparés en français parlé. *Recherches sur le Français parlé,* vol. (16) : 189-206.

Pallaud B. (2002a). Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le Français parlé,* vol. (17) : 79-102.

Pallaud B. (2003a). Erreurs d'écoute dans la transcription de données orales. In *Actes du colloque Transcription de la Parole normale et pathologique, revue Parole,* vol. (22-23-24) : 267-294.

Roubaud M.-N. (2003). Du bon usage des amorces dans la transcription de corpus. *Recherches sur le Français parlé,* vol. (18). A paraître.

Schegloff E., Jefferson H. et Sachs H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, vol. (53/2) : 351-382.

Shriberg E. (1999). Phonetic Consequences of Speech Disfluency.In *Symposium on The Phonetics of Spontaneous Speech, Proceedingsof the International Congress of Phonetic Sciences*, vol. (1) : 619-622.

Vers une recherche automatique des marqueurs de la segmentation du discours

Sophie Piérard[1], Liesbeth Degand[2], Yves Bestgen[3]

[1]UCL/PSOR –1348 Louvain-la-Neuve – Belgique – sophie.pierard@psp.ucl.ac.be
[2]FNRS – UCL/LIGE –1348 Louvain-la-Neuve – Belgique – degand@lige.ucl.ac.be
[3]FNRS – UCL/PSOR –1348 Louvain-la-Neuve – Belgique – yves.bestgen@psp.ucl.ac.be

Abstract

To study linguistic expressions signalling thematic breaks in large text corpora automatic procedures for the identification of these breaks are indispensable. In the present study, we test the effectiveness of four indices of cohesion whose calculation can be automated. We show that these indices make it possible to differentiate between three categories of temporal segmentation markers.

Résumé

Afin de pouvoir étudier dans de grands corpus de textes le fonctionnement d'expressions linguistiques qui signalent les ruptures thématiques, il est indispensable de disposer de procédures automatiques capables d'identifier ces ruptures. Dans la présente étude, nous testons l'efficacité de quatre indices de cohésion dont le calcul peut être automatisé. Nous montrons que ces indices permettent de différencier trois catégories de marqueurs temporels de la segmentation.

Mots-clés : marqueur de la structure, segmentation automatique, anaphore, analyse sémantique latente.

1. Introduction[1]

L'objectif de cette étude est de proposer et de tester des procédures qui, à terme, devraient permettre la recherche automatique des expressions linguistiques qui signalent un changement de thème dans un texte. Ce genre d'expressions retient depuis de nombreuses années l'attention des chercheurs et une longue liste de candidats à cette fonction a été proposée : (van Dijk, 1982 ; Virtanen, 1992 ; Longacre, 1979). Les arguments empiriques présentés pour les soutenir sont généralement issus de l'analyse d'un nombre très réduit de textes (Stark, 1988 ; Virtanen, 1992). L'origine de cette limitation est d'ordre pratique. L'étude des marqueurs de la structure d'un texte présuppose la connaissance de cette structure. Classiquement, celle-ci est obtenue par une analyse linguistique fine ou par le recours à des juges auxquels on demande d'indiquer les ruptures thématiques qu'ils perçoivent (Bestgen, 1992 ; Passonneau et Litman, 1997). La complexité et le coût de ces procédures manuelles rendent l'étude de grands corpus impraticable. Pour dépasser cette limitation, il est nécessaire de s'appuyer sur des techniques qui permettent de déterminer automatiquement les changements de thèmes. De telles techniques ont été développées durant ces dix dernières années dans le champ de l'analyse automatique du langage. Elles visent à identifier les ruptures thématiques les plus importantes en s'appuyant principalement sur l'analyse de la cohésion lexicale entre les

[1] Cette recherche a bénéficié du soutien de la Communauté française de Belgique – Actions de recherche concertées et du Fonds de la Recherche fondamentale collective (FRFC/FNRS).

phrases successives d'un texte. Nous proposons d'employer ces techniques pour valider la fonction de marqueur de segmentation d'une expression linguistique en montrant qu'elle est fréquemment associée à des ruptures thématiques.

Pour qu'une telle approche soit viable, il est toutefois nécessaire de montrer que les techniques de segmentation sont suffisamment efficaces. Or, depuis quelques années, les procédures de validation sont devenues de moins en moins exigeantes. Alors que les premières études confrontaient la segmentation automatique aux jugements de lecteurs, on procède actuellement en soumettant à l'algorithme une série de textes artificiellement concaténés et en mesurant sa capacité à retrouver les frontières des différents textes (Choi, 2000 ; Utiyama et Isahara, 2001).

L'objet de la présente étude est d'évaluer la faisabilité de l'approche décrite ci-dessus. Il s'agit donc de répondre à deux questions : peut-on calculer des indices de continuité/discontinuité suffisamment fiables pour étudier le fonctionnement des marqueurs de la segmentation et, si oui, est-il possible d'automatiser leur calcul ? Comme situation de test, nous avons choisi un ensemble de marqueurs étudiés dans de petits corpus de textes. Leur intérêt majeur est qu'ils forment une échelle allant du signalement d'une rupture importante à celui d'une continuité importante (Section 2). Afin de montrer si la validité de cette échelle peut être confirmée par l'analyse d'un grand corpus de textes, nous avons comparé quatre indices de cohésion qu'il est possible de calculer, au moins d'une manière approximative, par des procédures automatiques (Section 3). La section 4 présente les résultats d'une expérience réalisée sur la base d'un corpus de 5 000 000 de mots.

2. Marqueurs de continuité/discontinuité

Comme marqueurs de la segmentation, nous avons sélectionné un ensemble d'expressions temporelles, présentées ci-dessous, dont l'importance a été soulignée par de nombreux auteurs (par exemple, Segal *et al.*, 1991 ; Virtanen, 1992). Introduits en début de phrase, ces dispositifs peuvent être employés pour signaler la structure hiérarchique d'une narration comme l'a montré l'analyse d'un petit corpus de narrations de journées (Costermans et Bestgen, 1991). Les marqueurs d'ancrage (*vers dix heures*) qui font référence à un temps absolu, extérieur à la narration, signalent les ruptures les plus importantes. Les marqueurs d'enchaînement (*puis*, *ensuite*) qui indiquent la succession des actions à l'intérieur d'une narration signalent les ruptures intermédiaires alors que le connecteur *et* souligne la continuité discursive.

Continuité →			Discontinuité
Et	Puis	Vers neuf heures	
	Enchaîneur	Ancreur	

3. Indices de continuité/discontinuité

Les deux premiers indices de continuité/discontinuité s'inscrivent dans la ligne des travaux de Passoneau et Litman (1997) qui ont proposé de détecter les changements de thème dans un texte sur la base, entre autres, des ruptures dans les chaînes référentielles. Le premier indice est une mesure référentielle inspirée du modèle de Kintsch et van Dijk (1978). Ces auteurs ont élaboré un système de représentation de la signification d'un texte au moyen d'une liste structurée de propositions composées d'un prédicat et d'un ou de plusieurs arguments. Dans ce modèle, deux phrases sont liées si les propositions qui les composent contiennent des

arguments communs. En suivant Costermans (1998), nous définissons les arguments comme étant des noms ou des pronoms. Nous avons donc identifié dans chaque phrase cible, c'est-à-dire qui contient un marqueur, le nombre d'arguments présents dans au moins une des cinq phrases qui la précèdent ainsi que le nombre d'arguments neufs. S'il y a continuité thématique, le nombre d'arguments communs devrait être important. Dans le cas d'une rupture, les arguments neufs devraient être plus nombreux. Cet indice a été calculé par un juge et par une procédure automatique basée sur la comparaison des noms et pronoms présents dans la phrase cible et dans celles qui la précèdent. La limite majeure de cette procédure automatique est que deux arguments peuvent faire référence à la même entité en prenant des formes différentes comme *commerçant/marchand* ou encore *Robert/il.*

La seconde mesure ne prend en compte que les anaphores grammaticales. Selon l'expression de Moeschler « l'anaphore est une expression référentielle non autonome ». Sa présence est donc la trace d'une continuité thématique. Nous avons donc comptabilisé dans chaque phrase cible l'ensemble des pronoms et adjectifs anaphoriques (personnel, possessif et démonstratif) en distinguant les anaphores intra-phrase (*Pierre fit démarrer sa voiture*) des anaphores inter-phrases (*Pierre devait aller à l'aéroport. Sa voiture était en panne*) seules pertinentes ici. L'automatisation de cette procédure a simplement consisté en l'identification dans la phrase cible de l'ensemble des pronoms et adjectifs pouvant être anaphoriques. Les problèmes majeurs de cette implémentation résident en la distinction entre les anaphores intra et inter-phrases et du fait qu'un mot peut endosser plusieurs fonctions grammaticales : *le son de la cornemuse/ son chien*.

Les deux dernières mesures de continuité/discontinuité thématique ont été extraites par une procédure automatique basée sur l'Analyse Sémantique Latente (ASL : Landauer *et al.*, 1998). Cette technique vise à construire un espace sémantique de très grande dimension à partir de l'analyse statistique des cooccurrences dans un corpus de textes. Cet espace sémantique peut ensuite être employé pour mesurer la proximité sémantique entre des phrases au moyen de la mesure classique du cosinus (Bestgen *et al.*, 2003 ; Foltz *et al.*, 1998). Récemment, l'analyse sémantique latente a été utilisée par Choi *et al.* (2001) pour développer un algorithme de segmentation automatique plus efficace que plusieurs procédures classiques. Une présentation détaillée et une évaluation de l'efficacité de cet algorithme sont données par Bestgen (2004).

4. Expérience

L'objectif de cette expérience est de tester l'efficacité des indices de continuité/discontinuité décrits ci-dessus pour l'étude du fonctionnement des marqueurs de la segmentation en montrant que les marqueurs d'ancrage, d'enchaînage et le connecteur *et* sont associés à des niveaux de continuité/discontinuité thématiques différents. Pour deux de ces indices, nous comparons également les valeurs obtenues par une procédure automatique à celles déterminées par un juge.

4.1. *Corpus et extraction des marqueurs en contexte*

Le corpus est composé de textes littéraires (romans, nouvelles et contes) extraits principalement des bases ABU et Frantext. Il contient 5 000 000 de mots. Les textes ont été découpés en phrases et lemmatisés au moyen du programme *TreeTagger* de Schmid (1994).

Pour en extraire les expressions appartenant aux trois catégories de marqueurs, nous avons recherché toutes les occurrences en début de phrases des formes suivantes :

– « A midi », « Vers midi », « A minuit », « Vers minuit », « A *x* heure(s) », « Vers *x* heure(s) » où *x* représente un nombre compris entre un et vingt-quatre.
– Puis, Ensuite
– Et

Sur cette base, on a procédé à une sélection aléatoire d'un échantillon de 90 phrases pour chacune des trois catégories de marqueurs. Cette sélection a été effectuée en prenant en compte les œuvres littéraires et les auteurs de telle sorte qu'un nombre équivalent de phrases appartenant à chaque condition soit extrait d'un auteur donné. Enfin, on a extrait le contexte qui entoure les phrases contenant les marqueurs, soit les cinq phrases qui précèdent la phrase cible et les quatre phrases qui suivent celle-ci.

Pour l'estimation de la continuité/discontinuité par les chaînes référentielles, les deux indices suivants ont été calculés :
– Indice référentiel : le nombre d'éléments liés divisé par la somme du nombre d'éléments liés et d'éléments non liés.
– Indice anaphorique : le nombre total d'anaphores identifiées dans la phrase cible.

Pour les analyses manuelles de ces deux indices, un sous-échantillon de plus ou moins 50 segments par catégorie de marqueurs a été extrait. Dans chacun de ces segments, les marqueurs ont été supprimés par une personne qui n'a pas participé au codage manuel des phrases cibles. Un premier juge a calculé pour chaque extrait les deux indices de continuité définis ci-dessus. Afin d'estimer la fiabilité de cette évaluation, un second juge a analysé indépendamment 12 extraits de chaque type de marqueurs. Pour chacun des deux indices, la corrélation entre les évaluations des deux juges est au moins égale à 0.79.

Pour les analyses basées sur l'analyse sémantique latente, nous avons extrait du corpus initial un espace sémantique en respectant les paramètres proposés par Bestgen (2004). Le premier indice (Cosinus) employé est basé sur les cosinus entre une phrase cible (p) et celles qui la précède (p-1) et qui la suit (p+1). Si une phrase introduit un changement de thème, son cosinus avec celle qui la précède devrait être plus petit que celui avec la phrase qui la suit. Nous utiliserons donc comme indice : cos(p,p+1) - cos(p-1,p). Le second indice (Segment) est obtenu à partir de l'algorithme de segmentation de Choi *et al.* (2001). Celui-ci a été paramétré pour identifier d'une manière récursive les ruptures entre les 10 phrases d'un extrait en allant de la plus importante à la moins importante. Une rupture est donc d'autant plus forte qu'elle a été identifiée parmi les premières. L'indice employé est l'ordre d'importance de la segmentation qui précède la phrase cible en valeur relative, c'est-à-dire divisée par le nombre maximum de ruptures (9). Comme pour les autres indices, une valeur élevée indique à chaque fois une forte continuité.

4.2. *Résultats*

Comme indiqué ci-dessus, la procédure manuelle n'a été appliquée qu'au sous-échantillon de 152 phrases cibles. La procédure automatique a, par contre, été appliquée à ce sous-échantillon, mais aussi à l'échantillon total de 270 phrases cibles. Pour évaluer la fiabilité des analyses automatiques, nous avons corrélé les indices obtenus par ces procédures avec les évaluations faites par le juge. Tant pour l'indice référentiel que pour l'indice anaphorique, la corrélation entre est de 0.59 ($p<0.0001$). Au vu du caractère très rudimentaire de la procédure d'analyse automatique, ce résultat est très encourageant. Comme il n'y a pas de procédure manuelle pour calculer les indices obtenus à partir de l'ASL, ceux-ci n'ont été calculés que pour l'échantillon complet.

Le tableau présente les valeurs moyennes de continuité pour les trois catégories de marqueurs. La présence de différences statistiquement significatives a été déterminée par des analyses de variance à un critère.

	Indice référentiel			Indice anaphorique			ASL (Auto.)	
	Manuel** N=152	Auto* N=152	Auto** N=270	Manuel*** N=152	Auto*** N=152	Auto*** N=270	Cosinus* N=270	Segment** N=270
Et	0.44	0.39	0.35	1.00	1.63	1.60	+0.03	0.57
ENC	0.34	0.33	0.30	0.94	1.37	1.40	+0.02	0.55
ANC	0.23	0.26	0.24	0.22	0.61	0.60	-0.05	0.43

Valeurs moyennes de continuité pour les trois catégories de marqueurs.
Légende : Une valeur élevée indique une continuité forte. * $p<0.05$; ** $p<0.01$; *** $p<0.001$

Globalement, les valeurs observées pour les trois conditions sont ordonnées comme attendu. Le connecteur *Et* est associé à la plus grande continuité ; le marqueur d'ancrage à la plus forte discontinuité. Toutefois, pour trois des quatre mesures, le marqueur d'enchaînement est nettement plus proche de *Et* que de l'ancreur.

5. Conclusion

Même si cette étude présente plusieurs limitations (analyse d'un seul ensemble de marqueurs, caractères rudimentaires de certaines procédures d'estimation automatique des indices), ses implications nous semblent importantes à plus d'un titre. Tout d'abord, nous avons pu analyser dans un grand corpus de textes les fonctions de marqueurs de la continuité/discontinuité d'un ensemble d'expressions temporelles. Si les résultats confirment globalement l'échelle postulée, ils suggèrent la nécessité d'une analyse plus profonde de la distinction entre les enchaîneurs et le connecteur *Et*. De plus, cette méthodologie peut être employée pour comparer d'autres marqueurs. Il serait, par exemple, intéressant d'opposer les ancreurs à des expressions comme « *Une heure plus tard* » qui ne semblent pas remplir la même fonction dans le discours (Bestgen et Vonk, 2000 ; Zwaan, 1996). Plus généralement, la méthodologie décrite devrait permettre l'utilisation d'une approche plus heuristique : plutôt que de vérifier si une expression linguistique est associée à des ruptures importantes, on pourrait rechercher les ruptures les plus importantes afin de déterminer les expressions linguistiques qui y occurrent fréquemment.

Au niveau des techniques de segmentation automatique des textes, un résultat mérite d'être mis en exergue. L'indice anaphorique a donné lieu à des différences entre les trois catégories de marqueurs beaucoup plus significatives que celles obtenues avec les autres indices. Or, cet indice est rarement pris en compte par les algorithmes de segmentation automatique, qui, classiquement, commencent par supprimer les mots les plus fréquents et les mots outils. S'il est exact que ce genre de termes pose problème pour l'analyse de la cohésion lexicale, il serait intéressant de développer des approches qui combinent la cohésion lexicale et grammaticale.

Enfin, la présente étude s'inscrit dans le programme de recherches initié par Passoneau et Litman (1997). Ces auteurs ont proposé d'extraire la segmentation d'un texte sur la base de la cohésion lexicale, mais aussi de l'analyse des dispositifs linguistiques qui signalent la présence de changement de thème. Comme nous l'avons expliqué ci-dessus, nos indices référentiels et anaphoriques sont très proches de leur mesure des chaînes référentielles. Nos indices

présentent toutefois un avantage majeur : ils peuvent être, au moins partiellement, automatisés. Combiner en une seule procédure les indices référentiels, une analyse de la cohésion lexicale, par exemple par la procédure de Choi employée ici, et les marqueurs de la segmentation devrait permettre le développement d'un algorithme de segmentation particulièrement efficace.

Références

Bestgen Y. (2004). Analyse sémantique latente et segmentation automatique des textes. In *Actes des JADT 2004.*

Bestgen Y. et Vonk W. (2000). Temporal adverbials as segmentation markers in discourse comprehension. *Journal of Memory and Language*, vol. (42) : 74-87.

Bestgen Y., Degand L. et Spooren W. (2003). On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: an explorative study. In Lagerwerf L., Spooren W. et Degand L. (Eds), *Determination of information and tenor in texts.* Nodus Publikationen : 189-202.

Bestgen Y. (1992). Le textomètre : un outil pour l'étude de la structure du discours et de son marquage. *Psychologica Belgica*, vol. (32) : 141-167.

Choi F. (2000). Advances in domain independent linear text segmentation. In *Proceedings of AACL'00* : 26-33.

Choi F., Wiemer-Hastings P. et Moore J. (2001). Latent Semantic Analysis for Text Segmentation. *NAACL'01* : 109-117.

Costermans J. (1998). *Les activités cognitives : raisonnement, décision et résolution de problèmes*. De Boeck Université.

Costermans J. et Bestgen Y. (1991). The role of temporal markers in the segmentation of narrative discourse. *CPC: European Bulletin of Cognitive Psychology,* vol. (11) : 349-370.

Foltz P.W., Kintsch W. et Landauer T.K. (1998) The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, vol. (25) : 285-307.

Kintsch W. et Van Dijk T.A. (1978). Toward a model of text comprehension and production. *Psychological Review,* vol. (85/5) : 363-394.

Landauer T.K., Foltz P.W. et Laham D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, vol. (25) : 259-284.

Longacre R.E. (1979). The paragraph as a grammatical unit. In Givón T. (Ed.), *Syntax and Semantics,* vol. (12), *Discourse and Syntax*. Academic Press : 115-134.

Passonneau R.J. et Litman D.J. (1997). Discourse Segmentation by Human and Automated Means. *Computational Linguistics*, vol. (23) : 103-139.

Schmidt H. (1994). *Probabilistic Part-of-Speech Tagging Using Decision Trees.* Version électronique disponible sur [http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html].

Segal E.M., Duchan J.F. et Scott P.J. (1991). The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse Processes*, vol. (14) : 27-54.

Stark H.A. (1988). What do paragraph markings do? *Discourse Processes*, vol. (11) : 275-303.

Utiyama M. et Isahara H. (2001). A Statistical Model for Domain-Independent Text Segmentation. In *Proceedings of ACL'2001* : 491-498.

van Dijk T.A. (1982). Episodes as units of discourse analysis. In Tannen D. (Ed.), *Analysing discourse: text and talk*. Georgetown University Press : 177-195

Virtanen T. (1992). *Discourse functions of adverbial placement in English*. Åbo Akademi University Press.

Zwaan R.A. (1996). Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. (22) : 1196-1207.

Lexicométrie sur corpus étiquetés

Bénédicte Pincemin

CNRS, LLI – Université Paris 13
Av. J-B Clément – 93430 Villetaneuse – France
pincemin@lli.univ-paris13.fr

Abstract

Tagged corpus are now widely available, and are of great interest for textual and linguistic studies. Some lexicometric softwares have new versions to handle such corpus, but these don't give complete satisfaction yet. However, a clear and powerful model of text for lexicometric procedures has been formalized, as a string of *positions*; in each position one or several *types* are instanciated, from one or several sets of types, such as a set of spellings, or a set of lemmas, or a set of grammatical codes.
As regards the types definition, the way these kinds of linguistic information are recorded (the *record axes*) should not be confused with the views one can wish for a lexicometric analysis (the *analysis axes*). Actually, record axes are often irrelevant analysis axes. As regards the string of positions, some positions may be removed for the purposes of the analysis, so as to define the appropriate *background* retained from the *text*. Then the analysis can also be focussed on a given *pattern*, standing out against the background. We finally propose means to complete the results' display. These are naturally expressed and organized according to the analysis axis, but the introduction of views from some other axes may clarify, adjust or enrich their interpretation.

Résumé

Devant la disponibilité et l'intérêt des corpus étiquetés, l'adaptation des logiciels de lexicométrie n'est pas encore pleinement satisfaisante. À cependant été explicité un modèle lexicométrique du texte, comme suite de *positions* en chacune desquelles s'instancie un *type*, et ce éventuellement pour plusieurs familles de types (graphies, lemmes, codes grammaticaux par exemple).
Il convient d'abord de pouvoir redéfinir des *dimensions d'analyse* fixant les types considérés, qui ne sont pas directement la reprise des dimensions d'enregistrement des informations dans l'étiquetage, celles-ci étant généralement non pertinentes si elles sont utilisées telles quelles. Quant aux positions, il est intéressant de pouvoir en masquer certaines (*filtre fond / texte*) puis de focaliser les calculs sur un motif donné (*sélection forme / fond*). Enfin, bien que les résultats doivent rester présentés selon la dimension d'analyse qui les structure, nous proposons des manières de leur associer des éclairages selon d'autres dimensions, pour clarifier, nuancer ou enrichir l'interprétation.

Mots-clés : statistique textuelle, linguistique de corpus, lemmatisation, étiquetage morphosyntaxique, interface, Weblex, Hyperbase.

1. Avancées et limites des réalisations actuelles

Les corpus étiquetés se développent et deviennent les terrains d'analyse privilégiés pour les études textuelles et linguistiques. Le cas le plus fréquent d'étiquetage, que nous développerons ici, est celui de l'étiquetage morphosyntaxique, réalisable automatiquement sur de grands volumes de textes. Un tel étiquetage segmente le texte en mots (sur des critères plus linguistiques que le découpage par repérage de caractères délimiteurs) et associe typiquement à chaque mot son lemme, sa catégorie morphosyntaxique et ses informations de flexion.

Les logiciels de lexicométrie, initialement conçus pour traiter des corpus segmentés en graphies, se sont ouverts aux corpus étiquetés selon des voies relativement différentes.

L'approche la plus simple consiste... à ne rien changer. Les informations apportées par l'étiquetage peuvent être elles-mêmes considérées comme des « mots », et l'on peut faire des calculs lexicométriques sur le corpus écrit comme une succession de lemmes, ou encore écrit comme une succession de codes grammaticaux, ou tout simplement écrit comme la succession de ses mots. Les limites de cette approche sont évidentes : les calculs se placent dans l'une ou l'autre de ces vues, sans qu'aucun lien ne puisse être fait entre les occurrences se correspondant en fait d'une vue à l'autre. Les informations apportées par l'étiquetage sont perçues isolément, et chaque point de vue sur le corpus n'est réalisable que par le calcul (lourd) d'une nouvelle base, sans lien outillé avec les précédentes.

La seconde approche, illustrée par exemple par Hyperbase (Brunet, 2001 et 2002), consiste à intégrer ces multiples vues dans une même base. Le corpus est consultable avec son étiquetage complet. Pour les calculs généraux (vocabulaire présent avec ses fréquences, spécificités), l'affichage des résultats peut être donné en parallèle selon différentes vues intéressantes à considérer ensemble, par exemple graphies et lemmes. Mais en définitive, chaque calcul ne porte jamais que sur une vue du corpus, sans possibilité de croiser les différentes informations.

La troisième approche, développée par Weblex (Heiden, 2002), redéfinit la modélisation du corpus en lexicométrie pour la généraliser. Finalement, ce que l'on considère, c'est une suite de positions ; jusqu'à présent, à une position était associée une information (la graphie) ; avec les corpus étiquetés, c'est une série de propriétés qui peuvent être attachées à chaque position. Un objet d'étude peut alors être sélectionné par une équation combinant différentes propriétés : par exemple, on peut s'intéresser aux adverbes terminés en *-ment*, ou encore aux adjectifs qualificatifs suivant un verbe *être* non auxiliaire à la troisième personne. Ce modèle s'inspire du moteur de recherche mis au point par Christ (1994), et Weblex s'appuie sur cet outil. Les calculs lexicométriques apportés par Weblex restent cependant menés selon une vue à l'exclusion des autres (cf. paramètre *attribut d'occurrence implicite*) — à l'exception du calcul d'index. Une fois l'objet d'étude sélectionné, on considère toutes ses occurrences sous forme de graphies, ou de lemmes, ou de codes grammaticaux, tant pour les calculs que pour l'affichage.

Dans ces contextes, la pratique de la lexicométrie sur corpus étiquetés, enrichissante par la nouveauté des points de vue qu'elle permet, révèle néanmoins des limites importantes.

Tout d'abord, les analyses sur les codes grammaticaux sont peu exploitables à cause de la dispersion des informations, chaque propriété morphosyntaxique étant émiettée à proportion des autres propriétés auxquelles elle est associée au fil des occurrences. Il faudrait plutôt pouvoir considérer certaines propriétés et neutraliser certaines autres, en fonction des objectifs de l'analyse. Ainsi, Habert et Salem (1995), dans une analyse lexicométrique sur corpus étiqueté, en viennent à retravailler l'étiquetage soumis aux calculs, et concluent : « L'analyse des décomptes portant sur l'utilisation d'un système de catégories moins détaillé donne une image que nous avons jugé plus intéressante. »

De même, les autres objets d'étude (lemmes, graphies) ne sont pas si clairs ni satisfaisants qu'il paraît de prime abord (cf. § 2.2).

Enfin, le choix d'une dimension d'analyse (la graphie, le lemme, le code grammatical) s'avère être aussi celui de l'information affichée. Or pourquoi ne pas imaginer une concordance effectuée (et triée) sur les lemmes accompagnés des informations complètes de flexion, mais affichant également, pour plus de commodité, la graphie correspondante ? Ou encore, on pourrait sans extravagance vouloir le relevé du vocabulaire d'un corpus lemme par lemme, tout en détaillant pour chaque lemme les formes fléchies attestées dans le corpus.

Les approches actuelles semblent donc loin d'avoir définitivement résolu la prise en compte des corpus étiquetés dans les calculs lexicométriques. Montrant des perspectives extrêmement riches (comparaison des calculs selon différentes vues, définition d'objets d'étude par de multiples informations linguistiques complémentaires, généralisation du modèle lexicométrique) tout en frustrant l'utilisateur par les limites qui persistent, les réalisations actuelles sont motivantes et invitent à poursuivre la réflexion si bien lancée.

2. Modélisation

2.1. *Le modèle lexicométrique du texte*

Les calculs lexicométriques considèrent les textes comme une suite de positions, chacune de ces positions réalisant une occurrence d'une unité type. L'étiquetage s'intègre très bien à ce modèle en considérant que chaque position est multidimensionnelle, c'est-à-dire porteuse de plusieurs informations (graphie, lemme, code grammatical, etc.) (Heiden, 2002) qui sont elles mêmes autant d'occurrences renvoyant à des familles de types différentes (les types sont alors l'ensemble des graphies différentes, ou l'ensemble des lemmes, ou l'ensemble des codes grammaticaux). Cette modélisation ne rend évidemment pas compte de toute la réalité linguistique (où les délimitations d'unités ne sont pas toujours franches ni identiques d'un niveau d'analyse à un autre), mais elle offre néanmoins un terrain d'analyse fructueux.

2.2. *Dimensions de codage, dimensions élémentaires et dimensions d'analyse*

Ainsi, le texte n'est plus seulement une suite de mots, mais il est aussi, et même en même temps, une suite de lemmes ou une suite de codes grammaticaux. Cependant, chacune de ces dimensions prise indépendamment n'est pas toujours un bon objet d'étude au plan linguistique, et il faut donc se donner la possibilité de redéfinir des objets d'étude pertinents à partir des informations à disposition sans être limité à les reprendre telles quelles.

Prenons d'abord le cas des graphies : considérer les graphies isolément est parfois fait pour retrouver le cadre de travail d'avant l'étiquetage, pour évaluer, par comparaison, les changements apportés par l'étiquetage. Or les outils d'analyse linguistique peuvent être capables de reconnaître des locutions et des mots composés, ou du moins segmentent le texte sur des critères non purement typographiques (utilisation d'un dictionnaire, de listes de mots, de grammaires). Ne peuvent donc pleinement coexister une dimension de découpage par reconnaissance de caractères délimiteurs (Lebart et Salem, 1994 : § 2.1.2) et des dimensions de lemmes et de codes grammaticaux. Du coup, une analyse qui serait sur les graphies seules perd beaucoup de son intérêt, puisque cette vue n'est qu'une approximation du traitement par découpage simple et correspond moins bien à une réalité linguistique claire que les graphies désambiguïsées. Les adeptes de la non-lemmatisation[1] se retrouveront mieux en effet avec les graphies désambiguïsées, qui combinent la forme fléchie des mots et le code grammatical (il serait équivalent de donner le lemme, puisque toutes les informations de flexion sont précisées dans le code grammatical). Cette dimension d'analyse distingue tout ce qui morphosyntaxiquement peut être distingué, sans opérer de regroupements *a priori*, permettant de ne pas masquer des divergences de comportement statistique entre les différentes flexions d'un même lemme.

Les adeptes de la lemmatisation quant à eux ne devraient pas choisir de travailler sur l'étiquette « lemme », mais sur les lemmes désambiguïsés, c'est-à-dire où l'information lexicale

[1] Les arguments scientifiques développés par exemple dans (Geffroy *et al.*, 1974) ne sont pas périmés par la disponibilité actuelle de corpus étiquetés.

de la graphie du lemme est associée à l'information grammaticale de sa catégorie[2]. Car bien souvent, ce que l'on appelle le lemme n'est en fait que la graphie du lemme. Ainsi, on ne s'intéresse pas aux occurrences de *devoir* ou de *boucher*, mais à celles de *devoir (nom)* et de *devoir (verbe)*, de *boucher (nom)* et de *boucher (verbe)*.

Enfin, les codes grammaticaux ont rarement un intérêt à être exploités tels quels et pour eux-mêmes. Ils combinent de multiples informations qu'il sera généralement plus clair d'étudier séparément, voire de neutraliser quand les distinctions opérées ne semblent pas pertinentes pour l'étude. Habert et Salem (1995) choisissent par exemple, pour le dépouillement des réponses à une question ouverte (corpus *enfants*), de ne considérer les flexions que pour les verbes conjugués. Pour le français, les étiquetages morphosyntaxiques fournissent des informations de deux natures : catégorie et flexion. La catégorie précise les « parties du discours » (nom, verbe, etc.) affinées en sous-catégories quelquefois par des « traits » transversaux (indéfini, interrogatif, etc.) qu'il est intéressant de pouvoir considérer indépendamment d'une partie du discours particulière. Quant à l'information de flexion, elle n'est pas toujours présente (un certain nombre de catégories sont invariables) ; elle couvre les variations en genre, nombre, personne, et la conjugaison (mode et temps) pour les verbes.

Cette particularité du code grammatical, qui se décompose en fait en plusieurs informations structurées, pourrait théoriquement également advenir dans les dimensions plus lexicales de la graphie ou du lemme : pourquoi pas coder en effet des composants lexicaux comme les radicaux, préfixes et suffixes, les morphèmes, etc.

Il faudrait donc finalement distinguer :

- les *dimensions de codage* : typiquement graphie fléchie, graphie du lemme, code groupant catégorie et flexion. Elles sont héritées des usages en matière d'étiquetage. Il n'est pas certain qu'il faille en rendre compte pour l'utilisateur, puisque ce n'est pas *a priori* pertinent pour l'analyse lexicométrique.
- les *dimensions élémentaires* : par exemple morphèmes (si disponibles, ou sinon graphies des formes fléchies et graphies des lemmes), catégories, sous-catégories et traits, flexion en genre, nombre, personne et conjugaison (modes et temps). Elles sont souplement combinables pour la définition de dimensions d'analyse.
- des *dimensions d'analyse* : principalement graphie fléchie (ou du lemme) + catégorie + flexion, graphie du lemme + catégorie, tout ou partie des informations de catégorie ou/et de flexion éventuellement combinées avec des informations lexicales (lemme). À la différence des dimensions de codage et des dimensions élémentaires, en nombre limité fixé par le format d'enregistrement, les dimensions d'analyse ne sont pas énumérées mais construites par l'utilisateur en fonction de ses objectifs. Certaines (comme la graphie ou le lemme, désambiguïsés) sont évidemment plus classiques que d'autres et pourraient être proposées de façon prédéfinie.
- et éventuellement des *dimensions d'affichage* : elles se construisent comme des dimensions d'analyse, et ne se distinguent que par leur rôle (le calcul s'appuie sur la dimension d'analyse choisie, et les dimensions d'affichage ne font qu'apporter des éclairages complémentaires pour l'interprétation des résultats, cf. partie 2.3.3). L'affichage peut motiver la construction de dimensions qui lui sont propres (peu propices aux calculs mais intéres-

[2] La procédure d'intégration de corpus étiqueté de Weblex définit une telle propriété d'occurrence.

santes pour l'exploitation des résultats), telles que la graphie fléchie seule qui, en contexte (par exemple dans une concordance), est plus lisible et n'est pas ambiguë.

2.3. *Les trois moments de l'analyse lexicométrique*

Étant donné un corpus découpé en unités linguistiques étiquetées, le modèle lexicométrique du texte laisse donc trois paramètres à préciser. Deux sont nécessaires pour le lancement d'un calcul : d'une part, le domaine d'action du calcul, autrement dit l'ensemble des positions, des unités linguistiques, qui forment l'image du corpus pour le calcul (c'est ce que nous appellerons plus loin la délimitation du *fond* dans le *texte*) ; d'autre part, ce qui permet, en chaque position considérée, de se rapporter à un type dans un paradigme de types réalisables dans le corpus. Un troisième paramètre apparaît pour la présentation des résultats : si ceux-ci correspondent à une dimension donnée, l'affichage d'autres dimensions peut faciliter leur lecture ou éclairer leur interprétation.

2.3.1. *Les étapes de sélection : définition du fond dans le texte, focalisation sur des formes se détachant sur le fond*

Dans les logiciels actuels, le premier paramètre annoncé ci-dessus est peu disponible. Il n'apparaît peut-être que dans des options de découpage lors de la compilation d'un corpus. Les élagages et filtrages proposés ne délimitent pas le *fond* dans le *texte*, mais contribuent plutôt à sélectionner les *formes* sur le *fond*.

En effet, un corpus soumis à une analyse lexicométrique comporte en fait trois strates incluses l'une dans l'autre. La première est le *texte*, qui ancre l'ensemble des positions disponibles, se matérialisant typiquement par la suite complète des mots, des ponctuations, des signes typographiques. Les deux autres strates sont le *fond* sur lequel se profilent les *formes*[3]. Le fond sert au décompte des positions qui donne le contexte de référence dans les calculs (taille du corpus et des parties, domaine de décompte des fréquences). La strate des formes est l'objet d'étude, sélectionné parmi les éléments du fond. Dans les affichages d'occurrences en contexte, il est souhaitable en général d'afficher le texte complet (ne serait-ce que pour des questions de lisibilité), *a minima* le fond et les formes qui s'y détachent. Des effets typographiques devraient alors permettre de repérer ce qui relève de chacun des niveaux texte / fond / forme, par exemple le texte en plus petites polices ou en italiques, les formes en gras.

D'où deux filtres de sélection sur un corpus : l'un, souvent négatif, pour définir le *fond* à partir du *texte* ; l'autre, habituellement positif, et quelquefois doublé d'un discret filtre négatif d'« élagage » (classiquement mots grammaticaux, mots de moins de trois lettres, hapax, seuil sur le score statistique), pour désigner les *formes* à reconnaître sur le *fond*, auxquelles on va s'intéresser. Le premier filtre opère en amont des calculs, en influant sur la définition du fond qui sert de référence aux calculs il peut modifier, de façon « invisible », les scores associés aux formes. Alors que le second filtre opère en aval des calculs : il sert à montrer les résultats sur un sous-ensemble choisi par l'utilisateur d'un calcul qui par ailleurs aurait souvent pu être mené sur l'ensemble du fond sans autre incidence que de grossir la taille du calcul.

Préciser la forme que l'on souhaite étudier sur le fond est facultatif pour les calculs sélectifs, à savoir pour lesquels un score statistique est calculé et un seuil borne l'étendue des résultats. Faute de forme particulière indiquée, le calcul s'applique alors à tous les types de la dimen-

[3] Dans cette communication, nous utilisons le mot « graphie » pour désigner les formes fléchies, afin de réserver le mot « forme » à cette idée que se donner un objet d'étude en lexicométrie c'est finalement étudier le contraste de formes qui se détachent sur un fond.

sion d'analyse choisie ayant des occurrences dans le fond, puis le seuil limite l'affichage aux cas les plus remarquables. Pour d'autres procédures, telles que la recherche de contextes ou la construction de concordances, il n'y a pas de score statistique calculé : la sélection forme / fond est alors incontournable, et est d'ailleurs beaucoup plus spontanément abordée (on fait la concordance de tel mot, ou on recherche les contextes de tel lemme).

Si la forme est un motif complexe (sur plusieurs positions), elle peut être soumise en argument aux calculs de scores statistiques (tels que spécificités)[4] dans la mesure où l'on accepte de sortir du modèle statistique initial, et de faire l'approximation que, étant données les tailles du texte et des parties en nombre de positions, la forme peut être considérée comme ponctuelle.

Le travail sur un fond différent du texte n'est pas non plus anodin du point de vue des modèles statistiques. Dans certains cas, on pourra considérer que le fond reste très proche du texte et en reste une bonne approximation. Dans d'autres cas, il faudra plutôt pouvoir convenir, pour lancer les calculs statistiques, que la distribution des types sur le fond accepte les mêmes modèles et lois statistiques que ceux sur lesquels reposent les procédures utilisées.

2.3.2. *Définition de la granularité (rapport occurrences / type) par le choix d'une dimension d'analyse*

Le deuxième paramètre qui intervient dans les calculs consiste à définir, par une combinaison de dimensions élémentaires[5], la dimension d'analyse qui fixe les types auxquels rapporter les occurrences (cf. § 2.2) : par exemple, le tableau 1 ci-après montre, pour un extrait de phrase, les types reconnus selon différentes dimensions d'analyse. En particulier, c'est selon la dimension d'analyse que s'opèrent les tris. (Notons que lorsque l'on considère des formes « longues » (sur plusieurs positions), chaque position peut se voir attribuer une dimension d'analyse propre. Un calcul n'est donc pas toujours paramétré par une seule dimension d'analyse.)

Concrètement, si l'on considère les verbes (en tant que fond), on pourrait choisir de ne s'intéresser qu'aux temps verbaux (indépendamment des personnes), et étudier leur répartition au fil du texte ; ou encore, on pourrait centrer une analyse sur les personnes, sans les disperser sur les catégories qui les portent (possessifs, verbes, pronoms personnels).

	Dimensions d'analyse			
Extrait de phrase :	graphies désambiguïsées	catégories et sous-catégories grammaticales		flexion en nombre
Le	Le + Da-ms-d	Da	(déterminant article)	s
deuxième	deuxième + Ao-ms	A	(adjectif qualificatif)	s
paramètre	paramètre + Ncms	Nc	(nom commun)	s
qui	qui + Pr-ms--	Pr	(pronom relatif)	s

[4] Les calculs sur les suites de mots se pratiquent de fait déjà dans les logiciels lexicométriques.

[5] On indique un paradigme en donnant simplement les dimensions de variation pertinentes. Une interface graphique pour ce faire pourrait s'inspirer notamment du masque de saisie conçu pour Hyperbase (Brunet, 2002 : figure 9 : *Le choix d'un code grammatical*). Une autre voie pour définir une dimension d'analyse consisterait à définir autant de filtres en cascade que de types : si *condition 1* alors *type 1* sinon si *condition 2* alors *type 2*, etc. sinon *type reste* (hétérogène). Cette voie complémentaire est sans doute moins centrale que la combinaison de dimensions élémentaires, qui par nature sont ajustées à l'articulation linguistique des informations et construisent des paradigmes homogènes.

intervient	intervient + Vmip3s-	Vm	(verbe principal)	s
dans	dans + Sp	S	(préposition)	
les	les + Da-mp-d	Da	(déterminant article)	p
calculs	calculs + Ncmp	Nc	(nom commun)	p
Nombre de types :	8	6		2 + 1[6]

Tableau 1. Exemples de types reconnus selon différentes dimensions d'analyse

Dans les logiciels actuels, ce paramètre est encore souvent réduit à la sélection d'une dimension de codage : par exemple, le paramètre *attribut d'occurrence implicite* dans Weblex, qui rend déjà de grands services. Mais on n'a guère le choix de la granularité, qui est maximale : typiquement, si l'on travaille sur l'étiquetage morphosyntaxique, c'est chaque étiquette complète qui est constituée en type, avec une dispersion évidente et gênante des résultats. Deux amorces de définition souple de la granularité existent cependant. D'une part, pour le calcul d'*index* de Weblex, on peut « composer l'index avec les champs » de son choix, ce qui revient à sélectionner les dimensions actives pour définir les types. D'autre part, dans les *listes* d'Hyperbase, on peut grouper ponctuellement des items successifs ; ce nouvel objet peut être soumis au calcul des spécificités (histogramme) (et aux statistiques sur tableau, telles que analyse factorielle ou arborée). On a des procédés comparables de sélection groupée dans d'autres logiciels, par expression CQP pour les rafales dans Weblex, et par expression régulière dans les *types généralisés* de Lexico3.

2.3.3. *Affichages multidimensionnels : dimensions complémentaires et projections de résultats*

Le troisième paramètre, relatif à l'affichage, permet d'indiquer une ou plusieurs autres dimensions complémentaires à la dimension d'analyse. Il intervient de façon différente selon le type de résultats produit.

Le résultat peut être des *occurrences en contexte* : extraction de contextes, concordances, mais aussi surlignage en contexte des motifs sélectionnés par un calcul avec seuil (segments répétés, spécificités, rafales,...). L'affichage par défaut pourrait être le texte entier (non réduit au fond), avec indication, aux positions sélectionnées, des types reconnus, suivis des affichages complémentaires demandés. Il reste souhaitable (hors éventuelles considérations techniques de débit et de volume) qu'en tout point le texte soit cliquable, avec affichage complet des informations associées à la position considérée.

Le résultat peut autrement se présenter sous forme de *listes de types*. Dans ce cas, les affichages complémentaires peuvent se présenter comme une liste des valeurs prises par les occurrences de chacune des positions de réalisation de chaque type, pour chacune des dimensions souhaitées. Le choix d'un affichage complémentaire devrait donc préciser l'ordre souhaité

[6] Certaines dimensions d'analyse produisent des valeurs indéterminées pour certaines positions, par exemple ici les prépositions (et les conjonctions) n'ont pas de marque de nombre. On peut considérer que ces occurrences sont à considérer de façon groupée sous un type supplémentaire que l'on pourrait appeler *indéterminé*. Cependant il s'agit d'un type *a priori* hétérogène, et on préfèrera éviter autant que possible la formation de tels types par un filtrage *texte* / *fond* qui écarte de l'analyse les positions régulièrement non décrites par la dimension d'analyse.

pour l'énumération des attestations, qui pourrait être classiquement alphabétique (plus généralement canonique[7]) ou par fréquence décroissante.

Si les formes considérées s'étendent sur plusieurs positions (syntagmes, segments répétés, etc.), les affichages complémentaires seraient sans doute à donner position par position, sans préciser à ce niveau toute la combinatoire effectivement réalisée, pour ne pas surcharger les résultats. Pour en savoir plus, il est toujours possible de relancer un calcul en changeant de dimensions d'analyse.

Des procédures de projection / comparaison pourraient être disponibles, notamment pour visualiser la portée des choix qui ont une incidence sur le calcul : sélection texte / fond (qui fixe le « corpus de référence »), ou dimension d'analyse (par exemple lemmatisation ou non). L'utilisateur pourrait ainsi placer côte à côte plusieurs listes de résultats. Le clic sur un item d'une liste sélectionnerait dans les autres listes les items qui ont des occurrences communes dans le texte ou la partie. Un double clic sélectionnerait une liste entière ; le même principe de projection sur les autres listes permettrait de repérer les items originaux des autres listes, ceux qui n'ont pas de correspondants dans la liste initiale.

3. Conclusion

Face à la disponibilité et à l'intérêt des corpus étiqueté, l'adaptation des logiciels de lexicométrie conduit à généraliser la modélisation lexicométrique du corpus en « positions » et « dimensions » attachées à chacune de ces positions. Cette communication propose quelques éléments pour prolonger et développer une telle modélisation. La notion de *dimension d'analyse* est introduite pour donner accès à des objets d'étude non prédéfinis par les habitudes de codage et à la granularité réellement adaptée aux problématiques de recherche. La détermination des éléments linguistiques soumis au calcul lexicométrique (les *formes* à profiler sur un *fond*, éventuellement différent du *texte* complet) serait alors opérable par la définition de motifs, combinant des éléments correspondant à des positions unitaires dans le corpus. À chaque élément serait associé *(i)* son repérage (critères de sélection pour reconnaître les positions qui réalisent cet élément), *(ii)* sa dimension d'analyse (de quel type, dans quelle liste de types, est-il une occurrence), *(iii)* son affichage (outre la dimension d'analyse choisie en *(ii)*, qui détermine les calculs et articule la présentation des résultats, d'autres dimensions peuvent être demandées pour apporter un éclairage complémentaire).

Malgré son appareillage théorique central, l'exposé garde un caractère expérimental et prospectif. Il vise à défricher un terrain de réflexion et d'expérimentation sans prétendre fixer un cadre de référence stable. En effet, à ce stade de notre cheminement, ce sont la conception et éventuellement la réalisation de développements concrets qui seront maintenant les mieux à même de faire le tri entre les propositions fructueuses et celles impraticables, comme de préciser ce qui n'a été que trop vite défini. Un travail de conception d'interface est déjà amorcé, non moins passionnant que cette première élaboration du modèle théorique.

[7] Nous appelons *ordre canonique* un ordre d'énumération méthodique et mnémonique, associé aux dimensions élémentaires et aux dimensions d'analyse lors de leur construction. L'ordre canonique de codes grammaticaux peut être différent de l'ordre alphabétique : on voudra par exemple placer *singulier* avant *pluriel*.

Références

Brunet Ét. (2001). *Hyperbase, logiciel documentaire et statistique pour la création et l'exploitation de bases hypertextuelles. Manuel de référence.* Version 5.2 (mai 2001). Laboratoire Bases, corpus et langage, UFR Lettres, Université de Nice.

Brunet Ét. (2002). Le lemme comme on l'aime. In *Actes des JADT 2002* : 221-232.

Christ O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of COMPLEX'94 (3rd Conf. on Computational Lexicography and Text Research)* : 23-32.

Geffroy A., Lafon P. et Tournier M. (1974). L'indexation minimale. Plaidoyer pour une non-lemmatisation. ENS de Saint-Cloud. Communication au *Colloque sur l'Analyse des corpus linguistiques : Problèmes et méthodes de l'indexation minimale*, Strasbourg, 21-23 mai 1973.

Habert B. et Salem A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. *Traitement automatique des langues*, vol (36/1-2) : 249-275.

Heiden S. (2002). *Weblex. Manuel Utilisateur*. Version 4.1 (janvier 2002), Laboratoire ICAR, UMR 5191, ENS Lyon.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Lamalle C., Martinez W., Fleury S., Salem A., Fracchiolla B., Kuncova A. et Maisondieu A. (2003). *Lexico 3, Outils de statistique textuelle. Manuel d'utilisation.* Version 3.41 (février 2003). Laboratoire SYLED – CLA2T, Université de la Sorbonne nouvelle Paris 3.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Rajman M., Lecomte J. et Paroubek P. (1997). *Format de description lexicale pour le français – Partie 2 : description morpho-syntaxique. Technical report GRACE, GTR-3-2.1.*

Un pas en avant, un pas en arrière (Vingt-cinq ans d'allocutions radiodiffusées du roi Juan-Carlos d'Espagne)

Carmen Pineira-Tresmontant

Université Paris 8 (Saint-Denis)
cpineirat@aol.com

Abstract

The analysis of a range of chronological selected texts, made up of the King of Spain broadcast speeches he delivered during a period of 25 years, shows an alternation between the words *patria* and *sociedad* which seems to reveal a communication strategy suitable for the political context of Transitional then modernization period of Spain.

Résumé

L'analyse d'une série textuelle chronologique, composée de discours radiodiffusés du roi d'Espagne prononcés sur une période de 25 ans, montre une alternance des termes *patria* et *sociedad* qui semble révéler une stratégie de communication adaptée au contexte politique de la Transition puis de la modernisation de l'Espagne.

Mots-clefs : lexicométrie, discours politique, séries textuelles chronologiques.

1. Introduction

Installé sur le trône par les lois de succession imposées par Franco, Juan Carlos d'Espagne ouvre, dès son avènement en 1975, une transition démocratique qui aboutira trois ans plus tard à l'adoption d'une Constitution parlementaire. Par la suite, il défendra la jeune démocratie espagnole contre les tentatives de retour à l'ancien régime[1]. Il favorisera ensuite l'alternance en politique intérieure en désignant le socialiste Felipe González comme Premier Ministre.

Cette ouverture démocratique sera à l'origine d'un processus d'intégration de l'Espagne dans l'Union Européenne. En Espagne, le monarque ne gouverne pas, le pouvoir exécutif est aux mains d'un gouvernement issu de la majorité parlementaire et des représentants élus des différentes régions autonomes. Pourtant, aux yeux de ses contemporains, le roi a incarné la modernisation et la profonde transformation du pays. Constitutionnellement, il est le chef des armées.

Dans le cadre de ses différentes fonctions, le roi s'adresse fréquemment aux espagnols tantôt lors d'interventions ritualisées (Noël, Nouvel An, etc.), tantôt de façon plus protocolaire. Dans cette étude, nous avons tenté de mettre en évidence l'évolution chronologique de cette communication médiatique en analysant les discours du monarque tout au long de la période qui va de son avènement en 1975 jusqu'à nos jours.

[1] Tentative de coup d'Etat marquée par l'irruption du Colonel Tejero à la tête de ses troupes dans la salle des Cortes le 23 février 1981.

2. La série chronologique *Navidad*

Le corpus sélectionné pour cette étude est constitué des 25 discours adressés par le roi à l'ensemble des espagnols, chaque année à la même date, selon un mode particulièrement ritualisé au moment des fêtes de Noël. Ces discours réunis dans le corpus ***Navidad*** forment une série textuelle chronologique qui va de décembre 1975 à décembre 1999[2].

Les séries textuelles de ce type font l'objet de recherches de la part des spécialistes de l'étude quantitative des textes. On appelle :

> séries textuelles chronologiques ces corpus homogènes constitués par des textes produits dans des situations d'énonciation similaires, si possible par un même locuteur, individuel ou collectif, et présentant des caractéristiques lexicométriques comparables. […] Leur étalement dans le temps doit permettre de les comparer avec profit, de mettre en évidence ce qui varie au cours du temps (Salem, 1993).

Au plan lexicométrique, le corpus ***Navidad*** est caractérisé par une extrême cohérence puisqu'il répond aux différents critères qui caractérisent les séries textuelles chronologiques :

- Les 25 discours sont prononcés par un seul et même locuteur, dans des conditions d'énonciation identiques (allocution radio-télévisée depuis le Palais Royal, lecture d'un texte préparé à l'avance) ;
- Les discours sont prononcés à la même date pendant toute la période retenue qui couvre 25 années consécutives ;
- Le destinataire est le même : la communauté nationale espagnole ;
- Le thème est lui aussi largement motivé par la période de la présentation des vœux.

On trouve, ci-après, un extrait du discours de 1975 (début de la série) :

AL PUEBLO ESPAÑOL EN NAVIDAD, radiotelevisado, 24 de diciembre de 1975
En estas fiestas de nochebuena y Navidad en que las familias españolas acentúan su sentido entrañable y parece que quisiéramos ser mejores, me dirijo a todos para felicitaros las pascuas y desearos un año 1976 lleno de venturas y felicidad.
El año que finaliza nos ha dejado un sello de tristeza, que ha tenido como centro la enfermedad y la pérdida del que fue durante tantos años nuestro Generalísimo. El testamento dirigido al pueblo español es sin duda un documento histórico que refleja las enormes calidades humanas y los sentimientos llenos de patriotismo sobre los que quiso asentar toda su actuación al frente de nuestra nación ./.../

Tableau 1.Extrait de l'allocution radiodiffusée du 24 décembre 1975

La série chronologique est constituée de parties dont la longueur varie peu, comme on peut le vérifier sur la figure 1 ci-après :

[2] Signalons en outre l'existence d'une seconde série du même type constituée par des discours prononcés quelques jours plus tard, le 6 janvier, à l'occasion de la *Pascua Militar*, au cours de laquelle le roi forme ses vœux à l'intention des membres des Forces Armées. Deux dates fixes, deux publics aux caractéristiques dissemblables, deux contextes de communication politique séparés seulement par quelques jours. La comparaison de ces deux séries textuelles fait l'objet d'une étude parallèle.

Figure 1. Caractéristiques lexicométriques relatives aux 25 discours du corpus Navidad

Les discours sont de petite taille puisqu'ils comportent entre 500 et 2500 occurrences et de 200 à 800 formes. L'ensemble du corpus retenu compte ainsi **32 190** occurrences, **4 825** formes et **2 685** hapax. Les caractéristiques du corpus sont présentées en annexe 1.

3. Évolutions lexicales dans le corpus

Comme c'est souvent le cas dans les études quantitatives sur des corpus de ce type, les typologies réalisées à partir du tableau croisant formes et périodes (*formes X années*) mettent en évidence une forte évolution du vocabulaire liée au temps. L'analyse des correspondances réalisée à partir du tableau des formes les plus fréquentes dans les 25 discours montre (figure 2) que les textes qui correspondent aux années consécutives occupent sur le diagramme des positions proches[3]. Pour mettre en évidence les formes qui caractérisent les dernières parties du corpus, on effectue une analyse des spécificités portant sur ces périodes. On sélectionne de manière similaire (par l'analyse des spécificités du groupe de parties qui se trouve sur la droite du graphique) les formes les plus employées dans les premières années du corpus.

[3] L'analyse factorielle des correspondances produite a été réalisée à partir du tableau des 400 formes de fréquence supérieure à 10 dans l'ensemble du corpus.

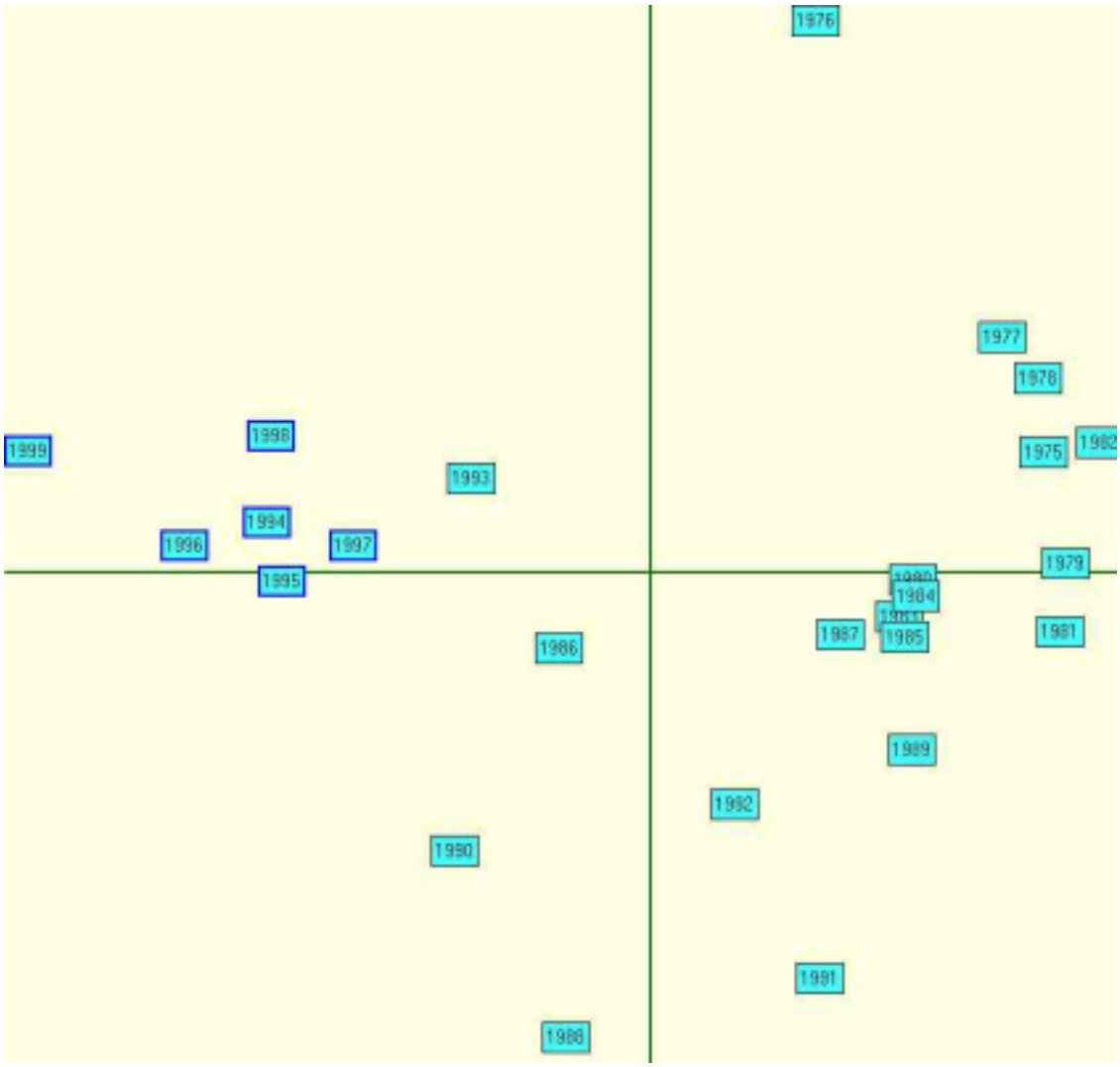


Figure 2.
AFC sur les parties du corpus Navidad
Les formes de fréquence > 10 sur le plan des deux premiers facteurs

Les changements les plus spectaculaires mis en évidence par ces analyses concernent les formes liées au morphème ***soci+*** (*sociedad*, *social*, etc.) dont l'utilisation va croître fortement à partir de 1993. Symétriquement, les occurrences de la forme *patria* tendent à disparaître, notamment à partir de l'année 1991.

De par leur évolution chronologique, les deux formes *patria* et *sociedad* dont l'évolution dans le corpus est mise en évidence par les méthodes quantitatives peuvent chacune prétendre à l'appellation de **prototype** pour des classes de formes dont l'évolution paraît particulièrement liée au temps. La première pour la classe des formes dont les occurrences ont tendance à disparaître au fil du temps, la seconde pour celles qui ont tendance à augmenter.

Bien évidemment ces évolutions ne sont pas uniformes, l'étude du rythme des modifications de leur fréquence fournit au contraire des enseignements intéressants. On peut voir sur la figure 3, la ventilation des formes *sociedad* et *patria* dans les 25 discours qui permet de repérer les moments forts de cette évolution. Bien entendu, les raisons de ces modifications dans la fréquence de l'emploi des formes sont à chercher dans l'histoire de l'évolution socio-historique de l'Espagne. Une certaine acception du terme *patria* appartient sans conteste au voca-

bulaire du passé franquiste. À l'inverse, l'utilisation de plus en plus fréquente des termes construits autour du morphème *soci+*, rend compte d'une prise en compte progressive des valeurs de la période moderne.

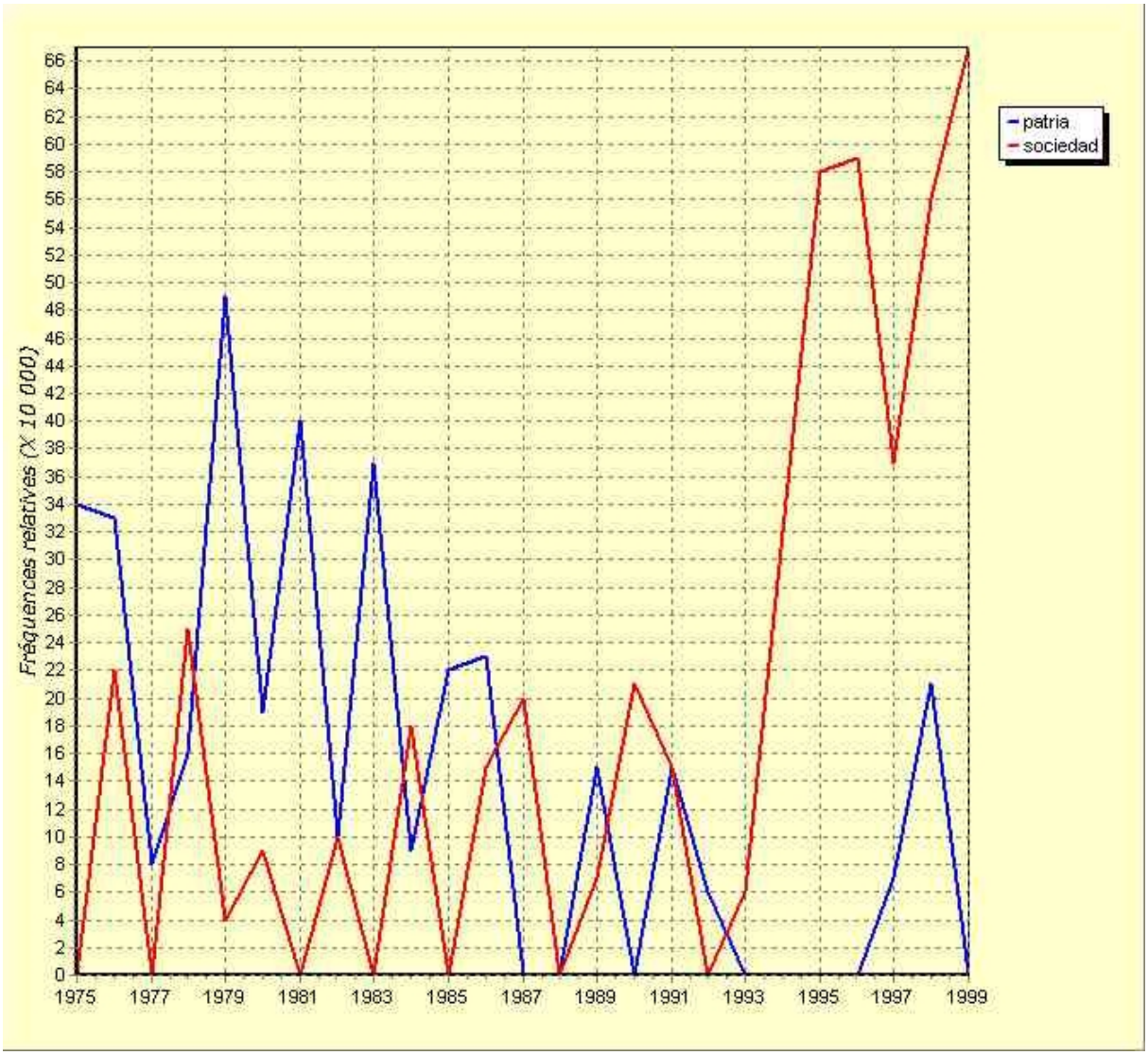


Figure 3.Ventilation des formes patria et sociedad par année

La forme *patria* représente la classe des formes plus particulièrement employées dans la première décade 1975-1985. Cette forme, totalement absente durant la période 93-96, trouve encore quelques emplois en 97-99. Dans l'ensemble, son évolution va dans le sens d'une utilisation moindre à mesure que l'on avance dans la période.

L'extrait de concordance que l'on peut voir au tableau 2 montre en outre que le contenu sémantique de cette forme s'est modifié au cours de la période couverte par le corpus. Les douze occurrences de *patria* attestées dans la partie qui correspond à l'année 1979 renvoient la plupart du temps à *la patria,* dans une acception largement usitée durant la période franquiste : *la patria* dont il est question ici est une entité qui réunit les valeurs et les élans de toute une communauté mais qui est, d'évidence, distincte des personnes humaines qui composent la communauté espagnole. Ainsi *la patria* n'est pas seulement une communauté regroupant les individus de nationalité espagnole. À cet égard la détermination est explicite : on

trouve une majorité de syntagmes nominaux *la patria* au début de la période, mais aussi *nuestra patria* et parfois *una patria* suivie d'une caractérisation (*sin fisuras, que constituya su aspiración, la patria y sus hijo, etc.).*

À partir de 1998, les usages de *patria* évoquent plus l'entité territoriale espagnole et non plus cette entité supérieure dotée d'une personnalité qu'était la *patria* des années 76-79 : *nuestra patria común, comunidades y territorios de nuestra patria, se encuentran fuera de nuestra patria.* Ainsi, la raréfaction de l'usage de *patria* correspond aussi au changement de son contenu sémantique et la perte de son signifié idéologique.

```
Partie : 1979, Nombre de contextes : 12
 del terruño viven y laboran por la patria , pensando en ella con nostalgi
es . § y ¿ qué menos que darles una patria sin fisuras , indestructiblemen
es de todas las regiones de nuestra patria - y hemos de sentir el orgullo
 ir más allá de las fronteras de la patria a buscar un sitio para vivir y
mpre el que dé ritmo al pulso de la patria . § a los intelectuales , que o
nción para estar a la altura que la patria exige , o vivir tranquilamente
oble corriente integradora entre la patria y sus hijos , entre éstos y la
poniéndoles un ideal de vida en una patria que constituya su aspiración su
ruir , de una vez para siempre , la patria que todos - todos nosotros , si
seremos capaces de hacer de nuestra patria un hogar del que se hayan deste
tria y sus hijos , entre éstos y la patria . § y todo resulta especialment
uel continente , que conviven en la patria común de un idioma que hablan e

Partie : 1998, Nombre de contextes : 3
anera generosa que españa , nuestra patria común , es de todos y para todo
omunidades y territorios de nuestra patria . § como he dicho en repetidas
 que se encuentran fuera de nuestra patria , a los que hago llegar un salu
```

Tableau 2-a.Extrait de la concordance de la forme : patria *année 1979 et 98*

À l'inverse, la forme *sociedad*, peu présente dans les premières parties du corpus, voit le nombre de ses occurrences croître notamment à partir de l'année 93. Ce terme, promis à une utilisation intensive, est d'abord introduit avec hésitation avant de connaître un essor notable à partir de 1994.

```
Partie : 1976, Nombre de contextes : 2
 que sacuden confuerza nuestra sociedad . con el firme asidero de nuestra
res más necesitados de nuestra sociedad , la creación de puestos de traba

Partie : 1978, Nombre de contextes : 3
 célula matriz y natural de la sociedad, y la que configura nuestro ser
parte en los azares de nuestra sociedad. para ellas - a las que , cada v
 en el perfeccionamiento de la sociedad, no va a torcerse .


Partie : 1999, Nombre de contextes : 9
nerales y en el conjunto de la sociedad española , creo que continua refl
l reconocimiento unánime de la sociedad hacia todas las víctimas del terr
remos e irrenunciables de toda sociedad civilizada , debemos reafirmar es
 ar ese compromiso común. § la sociedad vasca y el conjunto de la socieda
edad vasca y el conjunto de la sociedad española no van a permitir que el
os elementos esenciales de una sociedad que desea perfeccionarse en liber
 empleo estable , pues ninguna sociedad puede considerar que su desarroll
 españa aspira a construir una sociedad económica y socialmente avanzada
nuestras familias y de nuestra sociedad , prestar el respeto , el afecto
```

Tableau 2-b.Extrait de la concordance de la forme : sociedad, *années 1976, 1978 et 1999*

Ces quelques contextes permettent d'observer les deux usages principaux de la forme *sociedad* qui apparaissent au cours de la période considérée : *nuestra sociedad* constitue le pendant

sémantique de *nuestra patria*, regroupant l'ensemble des concitoyens espagnols d'une façon quasi-affective. Ce premier usage est prédominant au début de la période et se raréfie à partir de 1995. Le second usage constitue une description plus moderne de la société avec ses caractéristiques objectives notamment au plan économique et moins liée aux représentations subjectives des personnes qui la constituent (*España aspira a construir una* ***sociedad*** *económica y socialmente avanzada).*

4. Une évolution sous contrôle

L'observation la plus intéressante à propos de cette évolution conjointe vient du fait que ces changements ne se réalisent pas de manière progressive : dans la période qui s'étend de 1978 à 1985, chaque allocution semble corriger la timide évolution opérée lors de l'allocution précédente. La figure 3 montre cette évolution des fréquences pour les formes prototype *patria* et *sociedad*.

À l'accroissement de l'utilisation de la forme *sociedad*, très souvent accompagnée d'une sous-utilisation relative de la forme *patria*, que l'on observe grosso modo lors des années paires (76, 78, 80, 82, 84, 86), succède durant les années impaires une correction dans le sens inverse qui réactive la forme mise en sommeil et ainsi de suite jusqu'à la montée définitive de la forme *sociedad* qui survient en 95. Les spécificités de ces deux formes sont en opposition caractérisée au cours des deux sous-périodes 1977-1982 et 1986-1999 – (Annexe 2).

Une carte des sections (figure 4) réalisées à partir de la division du texte en paragraphes, représentés sous forme de carrés, permet encore d'affiner ces constats et de localiser des moments forts de l'évolution lexicale.

Sur cette figure produite par le logiciel Lexico3, à partir du corpus ***Navidad***, chacun des carrés représente un paragraphe du texte (portion de texte comprise entre deux retours chariot). Les carrés sombres signalent la présence de la forme *patria*, les carrés gris celles de la forme *sociedad*. On constate que les occurrences de la forme *patria* se raréfient au fil du temps.

On voit, sur cette figure, que les occurrences de *patria* dominent au cours des années 70 et 80, alors que celles de la forme *sociedad* voient leur fréquence augmenter dans les années 90, et ceci particulièrement à partir de 1993. Dans la première période, les apparitions de *sociedad* sont discrètes et relativement espacées : deux occurrences en 1976, aucune en 1977, trois en 1978, une en 1979, une en 1980, aucune en 1981, une en 1982, etc. Le caractère dominant de *patria* est moins marqué lorsqu'apparaissent les occurrences de la forme *sociedad* et d'avantage lorsqu'elles disparaissent ou sont peu nombreuses.

La forme *patria* apparaît deux fois en 1978 mais douze fois en 1979, deux fois en 1980 mais quatre fois en 1981, etc. On constate donc une sorte d'alternance d'une année sur l'autre dans l'emploi de *patria* et de *sociedad*, comme si la charge sémantique de ces deux formes avait quelque chose d'antinomique peu compatible dans le même discours. On avancera donc l'hypothèse que cette alternance peut être mise au compte d'un souci, voire d'une stratégie, d'introduire progressivement le vocabulaire nouveau, reflet et support de l'évolution idéologique.

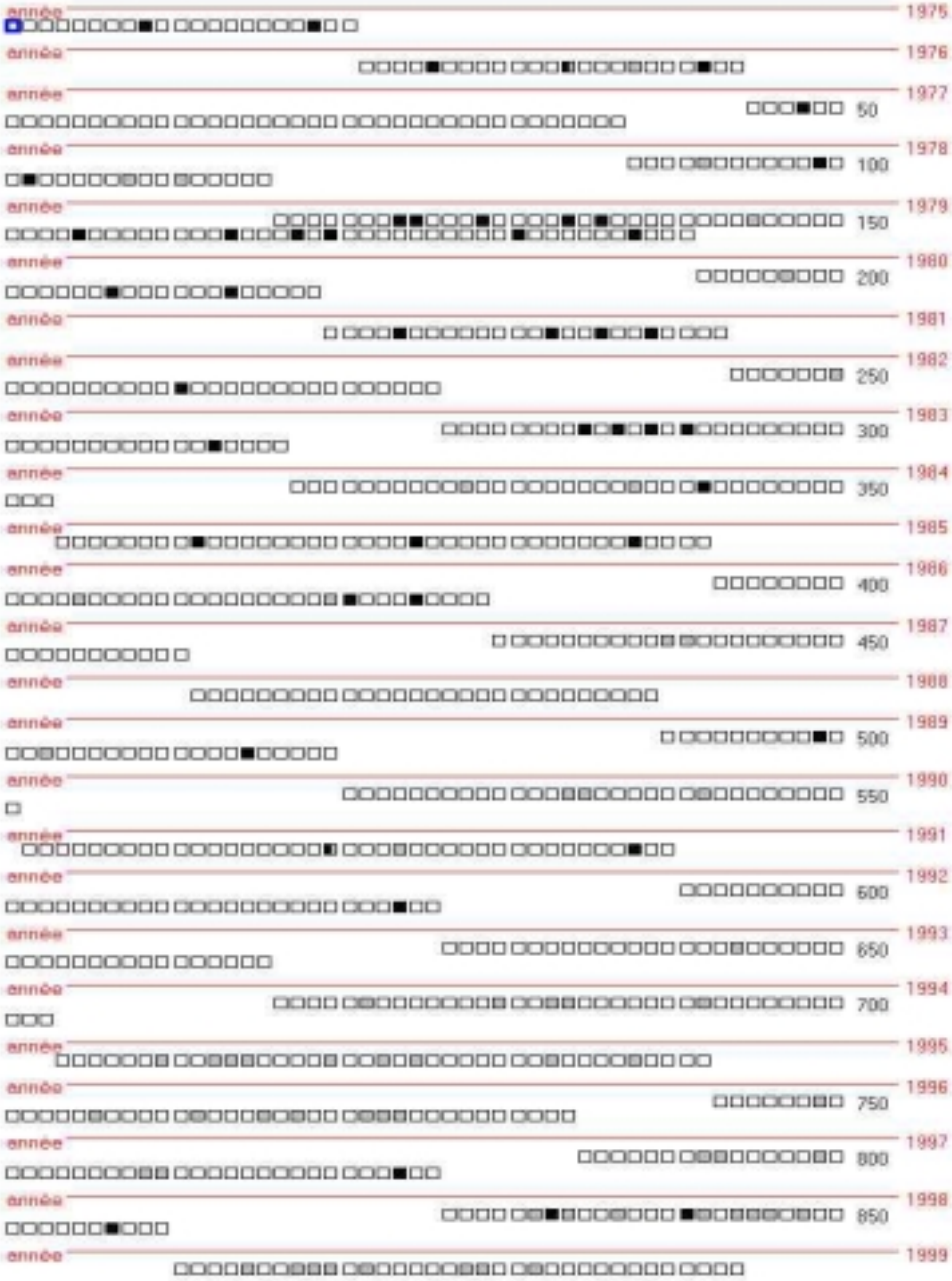


Figure 4. Carte des sections. Les forme patria ■ *et* sociedad

Sur la base de ces observations, on a sélectionné un corpus restreint aux années 78-85. Durant cette période on constate une alternance régulière dans l'usage des deux formes *patria* et *sociedad*. Les spécificités calculées pour les années paires 78, 80, 82, 84 qui présentent un usage intensif du terme *sociedad* mettent en évidence des formes dont les profils se rapprochent de celui de la forme *sociedad*. Les formes *esfuerzo, unidad, solidaridad, pueblo, futuro, voluntad, confianza, dificultades* présentent le même type de profil que la forme *sociedad*. On voit ainsi se préciser les contours d'un ensemble des formes qui deviendront dominantes au fil des années.

À l'inverse, les formes *mundo, europa* et le syntagme verbal *ha de* qui traduit le devoir apparaissent majoritairement dans les parties où domine la forme *patria*.

L'opposition entre les termes à spécificité positive et négative n'est pas seulement une opposition de termes mais aussi d'orientation de la communication : dans les parties où domine le terme *sociedad*, le roi souligne les efforts et les difficultés de la période en insistant sur les valeurs d'unité, de solidarité et de confiance, c'est un discours de volonté. À l'inverse sont délaissées les notions plus traditionnelles de patrie et de devoir.

Le même type d'analyse appliqué aux années impaires 79, 81, 83, 85 met en avant un grand nombre de termes liés à *patria* tels que *orgullo*, *bandera*, *ilusiones*, *desgracias* ainsi que des verbes tels que *pido*, *abandonemos*, *lamentemos*. L'analyse des spécificités permet de mieux cerner le sens donné par le roi à l'alternance des termes *sociedad* durant les années paires et *patria* au cours des années impaires : il y a alternance de signaux politiques qui diffèrent par leur contenu et les connotations qu'ils suscitent.

5. Un *nuestro* indicible en 1981

L'analyse des mutations chronologiques survenues au cours de l'ensemble de la période fait ressortir, parmi les résultats les plus marquants, la disparition de la forme *nuestro* dans le discours de 1981. Cette disparition est surprenante par comparaison avec l'usage fréquent de la forme *nuestro* dans les années qui précèdent mais aussi durant les années qui suivent. Quelques extraits des concordances de la forme *nuestro* (tableau 3) révèlent que l'emploi de cette forme correspond à une inflexion sémantique avant et après 1981.

del que fue durante tantos años	**nuestro**	generalísimo . el testamento dirigid
aje de amor que es la esencia de	**nuestro**	cristianismo , el cual nos exige sa
ejar paso a otros sentimientos ;	**nuestro**	pensamiento y nuestro corazón han d
imientos ; nuestro pensamiento y	**nuestro**	corazón han de pararse en aquellas
ante cuando se trata de servir a	**nuestro**	pueblo . soy consciente de las difi
la sociedad , y la que configura	**nuestro**	ser nacional - se armonizan y entie
oción mayor aún , el alcance de	**nuestro**	ideal . no es un pueblo fatigado el
ellos descansa la certeza de que	**nuestro**	camino en el perfeccionamiento de l
a inmensa capacidad luchadora de	**nuestro**	pueblo , que se crece cuanto más al
ealidad . § no abandonemos jamás	**nuestro**	orgullo español , nuestro ánimo dec
jamás nuestro orgullo español ,	**nuestro**	ánimo decidido , nuestra legítima p
nsecuencia de nuestra unión y de	**nuestro**	orgullo nacional . § yo os pido que
obusta constitución es médula de	**nuestro**	ser social ; el paisaje propio , en
ecen al escuchar fuera de españa	**nuestro**	himno nacional , al ver flotar nues
sumir , sin pesimismo ni temor ,	**nuestro**	futuro : un futuro digno de nosotro
" ? . § pongamos nuestra razón y	**nuestro**	corazón de españoles en la razón y
oñado alguna vez . § y si es así	**nuestro**	común anhelo de una españa mejor ,
a las de seguridad que protegen	**nuestro**	insobornable derecho a sentirnos es

Tableau 3-a. Extrait des concordances pour la forme nuestro *(années 75-80)*

l progreso y la modernización de	**nuestro**	país . § el problema del terrorismo
mencionar el creciente papel de	**nuestro**	país en el concierto internacional
a una mirada hacia atrás , hacia	**nuestro**	pasado , para que comprendamos hast
abilidad que hemos alcanzado con	**nuestro**	esfuerzo , debemos seguir ejercitan
irtudes que impulsaron con éxito	**nuestro**	proyecto nacional de convivencia en
arzo ha ratificado la solidez de	**nuestro**	proceso democrático y ha puesto de
vivimos momentos importantes de	**nuestro**	proyecto como nación , y os decía a
stamos construyendo el futuro de	**nuestro**	país . § el envejecimiento paulatin
estas aspiraciones profundas de	**nuestro**	pueblo

Tableau 3-b. Extrait des concordances pour la forme nuestro *(années 82-99)*

Comme le montrent les concordances ci-jointes, pendant la période des années 75-80, l'adjectif possessif *nuestro* est associé à des notions typiques du régime franquiste et, plus précisément des thématiques de la propagande franquiste. À travers le *nuestro*, le roi valorise

le christianisme (*nuestro cristianismo*), à l'essence nationale (*nuestro ser nacional*) et à l'orgueil espagnol (*nuestro orgullo español*), pour ne rien dire de l'évocation de feu *nuestro generalísimo*. Dès 1982, au contraire, le *nuestro* qualifie d'une façon plus neutre les désignants objectifs que sont le pays (*nuestro país*), l'effort (*nuestro esfuerzo*), le processus démocratique (*nuestro proceso democrático*) ou le concept original de projet comme nation (*nuestro proyecto como nación*).

Une explication de ce phénomène se présente immédiatement au niveau socio-politique à travers le contexte historique mouvementé de l'année 1981, au cours de laquelle le roi s'est opposé avec succès à une tentative de coup d'état. Le putsch de 81 rappelle l'existence d'une fracture entre les deux Espagne et semble priver le monarque de toute possibilité de référer à un corpus de valeurs communes. Après cette période mouvementée, le roi choisira d'utiliser le déictique *nuestro* pour qualifier des formes à forte connotation collective et orientée vers le changement et la modernisation.

6. Conclusion

Les méthodes lexicométriques ne permettent pas, à elles seules, de répondre à la question de savoir si l'évolution d'une série chronologique obéit à une stratégie de communication consciente, maitrisée et programmée ou résulte d'une suite de corrections successives résultant de réactions politiques à la réception des discours précédents. Cependant, au vu des résultats produits par les analyses lexicométriques, on peut avancer l'hypothèse que ce balancement dont la régularité est remarquable peut être mis au compte d'une volonté d'introduire progressivement un vocabulaire nouveau dont la suprématie deviendra manifeste par la suite, sans pour autant perdre le contact avec un auditoire traditionnel sensible à la permanence du vocabulaire du pouvoir. Ce constat sera le point de départ d'une recherche spécifique sur ce type d'évolution.

Annexe 1 :

Présentation générale du corpus

Année	Nb occurrences	Nb formes	Nb hapax	*Année*	Nb occurrences	Nb formes	Nb hapax
1975	584	295	226	*1989*	1317	561	422
1976	889	422	327	*1990*	1399	584	428
1977	1206	486	362	*1991*	1297	613	478
1978	1179	526	397	*1992*	1477	652	502
1979	2401	882	631	*1993*	1468	613	463
1980	1033	448	328	*1994*	1525	680	516
1981	996	456	340	*1995*	1706	717	536
1982	995	444	330	*1996*	1524	652	480
1983	1337	568	423	*1997*	1349	627	493
1984	1081	493	375	*1998*	1410	592	429
1985	1320	568	427	*1999*	1337	574	427
1986	1267	563	422	*1989*	1317	561	422
1987	984	472	367				

Annexe 2 :

***Spécificités des formes* Patria *et* Sociedad**

Années sélectionnées : 77 78 79 81 82

Forme	Frq. Tot.	Fréquence	Coeff.
Patria	34	17	4
sociedad	67	5	-3

Années sélectionnées : 86 90 94 95 96 97 98 99

Forme	Frq. Tot.	Fréquence	Coeff.
sociedad	67	51	11
Patria	34	1	-6

Références

Deroubaix J.-Cl. (2000). Le roi règne, il ne gouverne pas. Il s'adresse au peuple belge. In *Actes des JADT 2000*.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Salem A. (1988). Approches du temps lexical, statistique textuelle et séries chronologiques. *MOTS*, vol. (17).

Salem A. (1993). *Méthodes de la Statistique textuelle*. Thèse pour le doctorat d'État, Université de la Sorbonne nouvelle - Paris 3.

Tournier M. (2002). *Des sources du sens. Propos d'étymologie sociale*, vol. (3). ENS Éditions, Coll. Langages.

Contraintes syntaxiques et préférences sélectionnelles du verbe *entendre*

Sophie Piron
UQÀM
Département d'informatique et département de linguistique et de didactique des langues
CP 8888 Succ. Centre-Ville H3C 3 P8 Montréal – Québec – Canada
piron.sophie@courrier.uqam.ca

Abstract

This paper studies through a corpus analysis the way the French verb *entendre* is used. This corpus is made up of two language styles : one literary style and one neutral style. Only examples of the 20th century pertain to this corpus. A manual shallow labeling made it possible to group together the uses within a restricted set of syntactic patterns and to classify these patterns by the frequency they show in the corpus. The verb's meanings have also been listed on the basis of landmark dictionaries. Moreover, information on the semantic categories of verbal complements has been identified. The joint information given by syntactic patterns, verb's meanings and selectional preferences showed by the arguments allowed defining what guides the choice of the verb's meaning within a given context. The specificity of the analysis is due to the definition of syntactic patterns and semantic information based on the physiology of sensory perception.

Résumé

Cette étude s'attache à analyser le comportement du verbe français *entendre* en se basant sur une analyse de corpus. Ce corpus contient deux registres de langue : l'un littéraire, l'autre neutre. Il ne comprend que des emplois datant du XXe siècle. Un étiquetage manuel réduit a permis de regrouper les emplois du verbe étudié en un nombre restreint de patrons syntaxiques et de les classer selon leur fréquence d'apparition dans le corpus. Par ailleurs, une liste des sens du verbe a été établie sur la base de dictionnaires de référence. De plus, des informations sur la catégorie sémantique des compléments du verbe ont été identifiées. La mise en correspondance des patrons syntaxiques, des sens pris par le verbe et de la catégorisation des compléments du verbe *entendre* a permis de tracer les lignes directrices qui gèrent le choix du sens à donner au verbe *entendre* dans un contexte donné. La particularité de l'analyse est de reposer sur des patrons syntaxiques et sur des informations sémantiques issues de la physiologie de la perception sensorielle.

Mots-clés : verbe *entendre*, analyse de corpus, sous-catégorisation, contraintes syntaxiques, préférences sélectionnelles.

1. Problématique

Cet article s'attache à décrire le comportement du verbe *entendre* dans des emplois relevant de registres variés (à la fois littéraires et neutres). Il s'agit de découvrir des régularités syntaxiques et sémantiques qui permettraient de guider le choix du sens à donner au verbe *entendre* dans un contexte particulier.

2. Description du corpus

Cette analyse du verbe *entendre* est basée sur l'étude d'un corpus comportant deux registres de langue écrite : un registre neutre (dans notre cas, journalistique) et un registre littéraire. Le fait d'avoir deux registres permet d'obtenir une analyse à plus large couverture pour le comportement du verbe étudié. Cette analyse permettra également de contraster certains emplois

d'après leur registre d'appartenance. Le corpus littéraire est constitué de citations tirées du *Robert électronique* (1994). Seules les citations du XXe siècle ont été retenues. Le nombre de citations comprenant le verbe *entendre* s'élève à 720. Quant au corpus représentant le registre neutre, il a été construit grâce au programme de concordances en ligne *GlossaNet*. Un ensemble de 1360 citations comprenant le verbe *entendre* a ainsi été constitué. Ces citations sont issues du journal belge *La Meuse* pour une période allant du mois d'octobre 2001 au début du mois de janvier 2002. Ce journal représente un français qui n'est marqué que ponctuellement par des régionalismes[1]. Par ailleurs, ce journal donne accès à un français plus quotidien et plus neutre que le français du corpus littéraire. Le programme de concordances sélectionnait des citations avec 80 caractères avant et après le verbe *entendre.* Ces citations sont par conséquent souvent tronquées, mais elles ne le sont qu'en dehors de la proposition dans laquelle apparaît le verbe *entendre*. L'analyse n'est donc pas gênée par cet aspect.

Le corpus général (littéraire et neutre) comprend 2080 phrases employant le verbe *entendre.* Il a été encodé en format XML de manière à pouvoir procéder à plusieurs types de classements nécessaires à l'analyse. Il y a ainsi un certain nombre d'étiquettes qui encodent les caractéristiques de chaque citation (référence, analyse syntaxique, analyse sémantique, etc.). De plus, ce format permettra ultérieurement une consultation aisée du corpus.

3. Les sens du verbe *entendre*

Le verbe *entendre* appartient à une classe de verbes que la littérature nomme généralement *la classe des verbes de perception* (Levin, 1993). Nous parlerons plutôt des verbes de perception sensorielle, de manière à insister sur la spécificité de cette classe. De fait, les verbes qui en font partie permettent tous d'exprimer la perception au moyen d'un des cinq sens dont l'être humain dispose pour appréhender le monde extérieur : la vue, l'ouïe, l'odorat, le toucher et le goût. La classe de la perception sensorielle se subdivise ainsi en cinq champs, chacun consacré à une modalité sensorielle. Dans le champ de la perception auditive, le verbe *entendre* représente le sens générique de l'audition, par opposition à des verbes qui expriment également l'audition, mais avec des nuances sémantiques : *écouter*, *réentendre, ausculter,…*

Le but de cette étude étant de faire correspondre des significations avec des contraintes syntaxiques et des restrictions sélectionnelles, il fallait disposer initialement de la liste des sens du verbe *entendre.* Les notices de trois dictionnaires de référence (*Le Robert électronique* 1994, *Le grand usuel Larousse* 1997 et le *Trésor de la langue française* 1971-1994) ont été compilées. Par ailleurs, le travail sur le corpus a permis de mettre en lumière certaines nuances absentes des dictionnaires consultés. La liste des sens à laquelle nous avons ainsi abouti comprend 15 entrées qui se répartissent en deux domaines sémantiques (Piron, 2002) : le domaine sensoriel, qui relève de l'audition (7 entrées), et le domaine de l'attention mentale (8 entrées). Ces sens ont été intégrés aux annotations du corpus.

I. DOMAINE SENSORIEL		*Exemples*
1.	Compétence auditive	Je n'entends plus rien de l'oreille droite.
2.	Perception auditive	J'entends le son de la cloche.
3.	Attention auditive (écoute)	J'entends une conférence sur le verbe *entendre.*
4.	Écoute officielle	Les jurés entendent les témoins.
5.	Écoute favorable	Les revendications ont été entendues.

[1] Les régionalismes sont surtout le fait des discours rapportés en style direct.

6.	Mémoire auditive	On entend encore ses paroles.
7.	Transgression sensorielle	J'entends les couleurs.
II. DOMAINE DE L'ATTENTION MENTALE		*Exemples*
8.	Accepter	Étant entendu qu'il faut trouver une solution, …
9.	Juger	À entendre ses paroles, tout était perdu.
10.	Perception d'une situation	La multinationale ne l'entendait pas ainsi.
11.	Volonté	L'entraîneur entend conserver ce joueur.
12.	Relation – accord	Nous devrons nous entendre sur ce point.
13.	Relation réciproque	Ils sont faits pour s'entendre.
14.	Compréhension	Il a laissé entendre que tout était fini.
15.	Habileté	Il s'y entend pour troubler les esprits.

Tableau 1. Les sens du verbe entendre

4. Les patrons syntaxiques

Des patrons syntaxiques ont été établis pour classer les emplois du verbe *entendre* dans le corpus mis sur pied. Ces patrons encodent la sous-catégorisation syntaxique du verbe et, par conséquent, s'attachent uniquement à la complémentation essentielle. Le sujet n'est donc pas pris en compte à cette étape de l'analyse. Ce que le patron syntaxique encode, c'est si le verbe est employé avec un complément de type syntagme nominal ou s'il est utilisé avec un syntagme prépositionnel essentiel, avec une complétive infinitive, etc. Pris tels quels, les types de constructions syntaxiques sont nombreux. Nous en avons relevé une bonne soixantaine, mais, après analyse, ils se réduisent finalement à 9 patrons. Ainsi regroupés, ils permettent d'étudier le comportement syntaxique du verbe. Ces patrons apparaissent dans le corpus sous la forme d'annotations. Le corpus présente ainsi un étiquetage syntaxique partiel.

4.1. *Patron [*V* SN]*

Le verbe *entendre* est indiqué par le sigle *V* et son complément est de type SN, c'est-à-dire un syntagme nominal. Cela dit, ce patron regroupe non seulement les emplois du verbe avec un syntagme nominal objet (exemple 1), mais également des emplois dans lesquels l'objet est exprimé par un pronom (exemple 2) ou un clitique (exemple 3).

Exemple 1. emaine, depuis l'ouverture de leur salle, nous ne savons plus dormir. [Nous n'entendons pas la musique mais les basses.] Et quand ce n'est pas ça, c'est par (entendre_LaMeuse_241001_09)

Exemple 2. Avec une question simple : où en est la santé financière du club? [Comme on a entendu tout et son contraire,] nous voulions faire le point avec la personne la (entendre_LaMeuse_261101_09)

Exemple 3. chaque année. Les premiers jours les camions passent. [Puis après, on les entend le matin] mais ils ne passent plus par après. On circule avec peine. (entendre_LaMeuse_040102_10)

4.2. *Patron [*V* [V]]*

Ce patron identifie tous les emplois du verbe dans lesquels *entendre* est le verbe principal et est suivi d'une complétive infinitive dont le premier terme apparaissant dans la citation est le verbe de la complétive.

Exemple 4. le de Namur, propriétaire du site, en décidera. [Mais les anciens combattants entendent demander à ce que l'éventuel nouveau sanglier soit en pierre et scellé] (entendre_LaMeuse_031101_14)

Exemple 5. up de balai dans son train-train quotidien. L'occasion nous est offerte [de l'entendre disserter sur l'art de vivre en couple,] de fuir les habitudes, de tirer (entendre_LaMeuse_141201_06)

4.3. *Patron [*V* [x V]]*

Ce patron est très semblable au précédent : il identifie, comme lui, des complétives infinitives à la suite du verbe *entendre.* Par contre, cette fois, un élément intervient au début de la complétive infinitive (exemples 6 et 7). Le but d'une telle distinction est de déterminer si toutes les complétives infinitives présentent le même comportement sémantique ou si la présence d'un élément en position initiale dans la complétive infinitive a une influence quelconque sur le sens associé à cette structure syntaxique.

Exemple 6. Le général de Gaulle ordonnait l'action selon un « grand dessein » variable, puisqu'il était limité par le possible, variable lui aussi. [Il entendait l'accomplir par tous les moyens dont il disposait.] MALRAUX, Antimémoires, Folio, p. 157. (entendre_GrandRobert_cit20_492)

Exemple 7. [On entend les insectes siffler partout, parce qu'il y a des sources d'eau minuscules entre les roches, et de grands puits cachés dans les avens où l'eau froide attend.] J.-M. G. LE CLÉZIO, Désert, p. 101. (entendre_GrandRobert_cit20_564)

4.4. *Patron [*V* Compl]*

Ce patron représente tous les cas pour lesquels le verbe *entendre* est suivi d'une complétive conjuguée cette fois.

Exemple 8. (Il) fit pivoter sa main, paume dessus, paume dessous, [pour laisser entendre que le salaire était variable.] G. DUHAMEL, Salavin, VI, XV. (entendre_GrandRobert_cit20_183)

4.5. *Patron [*V* P]*

Ce patron identifie les cas dans lesquels le verbe *entendre* est suivi d'une phrase en discours direct, le P symbolisant la phrase citée.

Exemple 9. sujet était moins esthétisant. [« Des films de cette qualité, on en redemande » entendait-on à la sortie de la projection.] Des tickets pour des jeunes rouviro (entendre_LaMeuse_131101_09)

4.6. *Patron [AuxPass *V*]*

Ce patron comprend tous les cas d'emploi du verbe *entendre* avec l'auxiliaire *être*.

Exemple 10. - Oh! pas d'histoire, affirma-t-il. [L'affaire est entendue.] Francis CARCO, Jésus-la-Caille, III, 5. (entendre_GrandRobert_cit20_053)

4.7. *Patron [CLse *V*]*

Il s'agit des cas dans lesquels le verbe *entendre* est utilisé avec le clitique *se*. Des éléments peuvent être utilisés pour compléter la structure. Ainsi, par exemple, l'auxiliaire modal *faire* :

Exemple 11. i émerge finalement 24-26. [Dès l'entame du 5ème set, les supporters se font entendre.] Arabis paraît un peu moins stressé et se détache à 08-05. Le Naj (entendre_LaMeuse_291101_13)

4.8. *Patron [*V*]*

Dans ce cas, le verbe est utilisé seul, sans compléments essentiels.

Exemple 12. Notre vue n'est pas devenue plus perçante ni notre ouïe plus fine, mais nous avons créé des instruments [qui nous permettent de voir et d'entendre très au-delà de ce qu'auraient pu rêver les plus audacieux futurologues]. Emmanuel BERL, le Virage, 1972, p. 93 (entendre_GrandRobert_cit20_311)

4.9. *Patron [CLind *V*]*

Ici, le verbe est utilisé essentiellement avec le clitique indirect.

Exemple 13. Je discutais avec l'adjudant du bataillon les termes mêmes de son rapport, lui disant [qu'il n'y entendait rien,] que les accusations qu'il portait ne rimaient à rien du tout [...]. B. CENDRARS, la Main coupée, in OE. compl., t. X, p. 136. (entendre_GrandRobert_cit20_136)

5. La fréquence des patrons syntaxiques

La fréquence des patrons syntaxiques est utile à analyser dans la mesure où elle peut déboucher sur de fortes différences d'utilisation d'un patron à l'autre. Une telle information s'avère intéressante à encoder dans des systèmes de TAL. Dans le cas du verbe *entendre,* les neuf patrons syntaxiques qui ont été relevés se répartissent équitablement au sein des deux corpus. Tout d'abord, de manière générale, on constate que la construction de type [*V* SN] est, de loin, la plus fréquente. Les autres constructions que l'on retrouve le plus souvent sont [*V* [V]], ensuite [AuxPass *V*] et enfin [CLse *V*]. Les constructions [*V* [x V]] et [*V* Compl] sont beaucoup moins représentées. Enfin, les constructions [*V*], [*V* P] et [CLind *V*] sont sous-représentées.

Patron syntaxique	***La Meuse***		***Le Robert***		**Corpus général**	
V SN	554	40,73 %	378	52,50 %	932	44,81 %
V [V]	316	23,24 %	158	21,94 %	474	22,79 %
AuxPass *V*	182	13,38 %	35	4,86 %	217	10,43 %
CLse *V*	127	9,34 %	55	7,64 %	182	8,75 %
V [x V]	68	5,00 %	40	5,56 %	108	5,19 %
V Compl	73	5,37 %	30	4,17 %	103	4,95 %
V	18	1,32 %	20	2,78 %	38	1,83 %
V P	22	1,62 %	3	0,42 %	25	1,20 %
CLind *V*	0	0,00 %	1	0,14 %	1	0,05 %
total	1360	100,00 %	720	100,00 %	**2080**	**100 %**

Tableau 2. La fréquence des patrons syntaxiques du verbe entendre

Quelques différences apparaissent entre les deux corpus en ce qui concerne d'abord la construction [*V* SN]. Elle est davantage représentée dans le corpus littéraire avec 52,36 % (comparativement à 40,59 % dans le corpus journalistique). À l'inverse, le corpus issu du journal *La Meuse* utilise presque trois fois plus la construction [AuxPass *V*]. Il l'utilise même davantage que la construction [CLse *V*], alors que l'inverse se produit dans le corpus littéraire.

6. Les préférences sélectionnelles

De manière à guider le choix du sens à donner à une construction, des critères syntaxiques sont bien sûr essentiels, mais il faut aussi tenir compte de ce que l'on nomme *les préférences sélectionnelles* ou *les restrictions sélectionnelles* (Manning et Schütze, 2000). Il s'agit des types sémantiques qui caractérisent les arguments d'un verbe. L'exemple le plus souvent cité est le cas du verbe *manger*, qui se construit avec des arguments relevant du domaine de la nourriture. On parle de *préférences*, parce que ce ne sont là nullement des règles (pensons aux emplois métaphoriques qu'un verbe peut présenter).

Nous avons intégré à notre corpus des annotations qui identifient l'information sélectionnelle des arguments du verbe *entendre.* Des travaux dans ce domaine se sont basés sur WordNet ou GermaNet pour identifier les préférences sélectionnelles (Resnik, 1997 ; Clark et Weir, 2002).

Par rapport à ces travaux, la particularité des informations que nous avons insérées est d'être basées sur la physiologie de la perception auditive. Ce type de perception distingue trois modes auditifs (Mann et Liberman, 1983 ; Perez, 1994) : environnemental, phonétique et musical. Ces distinctions ont été appliquées à la catégorisation des compléments du verbe *entendre* et se sont avérées très utiles pour notre analyse.

7. Contraintes syntaxiques, préférences sélectionnelles et sens du verbe

Nous n'étudierons ici que les six patrons syntaxiques les plus représentés.

7.1. *Le patron [*V* SN]*

Cette construction est la plus fréquente, avec 44,81 % d'exemples du corpus. C'est en même temps une construction très polysémique puisqu'elle comptabilise 12 sens différents dans les deux corpus. Le tableau ci-dessous présente la répartition des sens correspondant à la construction [*V* SN] d'abord dans le corpus *La Meuse*, ensuite dans le corpus du *Robert électronique* et, enfin, dans les deux corpus pris conjointement. Bien évidemment, la répartition entre ces sens est très inégale. Sur l'ensemble des deux corpus, tous les sens relevant du domaine de l'audition (les 7 premiers sens du tableau ci-dessous) atteignent une représentation de 78,65 %. Et à l'intérieur de ce domaine, c'est la perception auditive qui correspond le plus souvent à la construction [*V* SN]. Le corpus littéraire fait d'ailleurs un usage très important de cette construction avec le sens de la perception auditive (75,40 %). Le corpus journalistique montre le même attrait pour la correspondance entre la construction [*V* SN] et le sens de la perception auditive (47,47 %), mais il montre surtout davantage de répartition des sens du domaine auditif en faisant un usage, plus restreint il est vrai, des sens d'écoute - attention (17,33 %) et d'écoute officielle (11,73 %).

Sens de *entendre*	***La Meuse***		***Le Robert***		**Corpus général**	
perception auditive	263	47,47 %	285	75,40 %	548	58,80 %
écoute – attention	96	17,33 %	13	3,44 %	109	11,70 %
écoute officielle	65	11,73 %	2	0,53 %	67	7,19 %
écoute favorable	4	0,72 %	0	0,00 %	4	0,43 %
compétence auditive	1	0,18 %	1	0,26 %	2	0,21 %
transgression modale	1	0,18 %	1	0,26 %	2	0,21 %
mémoire auditive	0	0,00 %	1	0,26 %	1	0,11 %
Comprendre	53	9,57 %	61	16,14 %	114	12,23 %
Accepter	21	3,79 %	10	2,65 %	31	3,33 %
perception – situation	29	5,23 %	2	0,53 %	31	3,33 %
juger	11	1,99 %	2	0,53 %	13	1,39 %
vouloir	10	1,81 %	0	0,00 %	10	1,07 %
total	554	100,00 %	378	100,00 %	932	100,00 %

Tableau 3. Les sens du verbe entendre *avec le patron syntaxique [*V* SN]*

En ce qui concerne les sens relevant du domaine de l'attention mentale (21,35 % des exemples), c'est la compréhension qui correspond le plus souvent à la construction [*V* SN]. Le corpus littéraire présente une forte préférence pour ce sens, même si d'autres sens sont associés à cette construction. Il s'agit essentiellement du sens de *accepter* qui apparaît aux côtés de la compréhension. Le corpus journalistique fait également un usage préférentiel de la construction transitive directe pour exprimer la compréhension, mais il montre surtout une

plus grande répartition d'usage parmi les sens utilisés : *perception – situation, accepter*, *juger* et *vouloir* (en ordre décroissant d'utilisation).

Il nous est impossible de présenter ici toutes les distinctions qui concourent à déterminer le sens associé à la construction [*V* SN]. Par contre, nous insisterons sur les éléments les plus saillants que notre analyse a permis de faire ressortir. En premier lieu, la caractérisation sémantique du SN qui est complément du verbe *entendre* permet de définir les grands choix de sens opérés. Un SN de type *musique* ou *source_musique* (7,53 % du corpus général, avec la construction [*V* SN]) ne permet, dans le corpus étudié, que deux interprétations relevant d'ailleurs du domaine de l'audition : la perception auditive ou l'écoute. Un SN de type *parole* ou *source_parole* (39,07 % du corpus général) peut prendre de nombreux sens, mais une très nette préférence va vers la perception auditive (60,88 % des SN de ce type prennent ce sens). Les autres sens associés à cette construction sont *l'écoute – attention, l'écoute officielle* et, dans les citations littéraires, le sens de *comprendre*. On remarque également que, avec le sens de la perception auditive, le corpus *La Meuse* préfère vraiment utiliser un SN_parole plutôt qu'un SN_source_parole, alors que cette préférence, si elle existe clairement dans le corpus littéraire, y est moins forte. Un SN de type *son* ou *source_son* (26,91 % du corpus général) montre une nette préférence pour le sens de la perception auditive (98,4 % des SN de ce type prennent ce sens). Les quelques rares autres cas relèvent de toute façon du domaine général de l'audition.

Il est intéressant de souligner un dernier point dans l'analyse de la construction [*V* SN] : il s'agit de l'expression de la compréhension. Dans ce cas-ci, l'analyse repose tantôt sur des critères de catégorisation sémantique du complément du verbe, tantôt sur des spécificités de la construction syntaxique. Les deux types de corpus montrent une préférence pour la construction « syntagme prépositionnel (*par*) – verbe – SN ». En voici un exemple :

Exemple 14. llissement de la population et l'augmentation de l'obésité. [Par diabète, on entend souvent « diabète sucré », qui se définit par un taux de sucre dans le sang] (entendre_LaMeuse_201001_33)

Dans le corpus neutre, 41,67 % des emplois de *entendre* au sens de *comprendre* avec le patron général [*V* SN] utilisent cette construction plus spécifique. Le chiffre monte à 55,76 % dans le corpus littéraire. Ce dernier utilise également la construction [*V* SN] avec un SN de type *parole* ou *hors_classe* pour signifier la compréhension. Ce procédé est délaissé par le registre neutre au profit d'un autre procédé (syntaxique) : l'utilisation de l'impératif (exemple 15).

Exemple 15. oline soit bien mise» dit-elle au directeur de l'établissement. Caroline? [Entendez sa perruque!] Des petites phrases en wallon, un dialogue quotidien, (entendre_LaMeuse_071101_09)

7.2. *Le patron [*V* [V]]*

La construction de *entendre* directement suivi du verbe de la complétive infinitive représente 22,79 % de l'ensemble du corpus. Si elle est associée à cinq sens du verbe *entendre*, elle n'est véritablement polysémique qu'entre deux de ces sens : la volonté et la perception. Il faut relever que ces deux sens relèvent chacun de domaines différents : perception et attention mentale. Comme on peut le constater dans le tableau ci-dessous, le corpus littéraire montre une large préférence pour l'association entre la construction [*V* [V]] et le sens de la perception auditive avec 81,65 % tandis que l'expression de la perception au moyen de cette même construction n'est que –comparativement– de 17,72 %. Le corpus journalistique présente un comportement différent : la construction est davantage utilisée avec le sens de la volonté (56,65 %), mais le sens de la perception est presque en concurrence avec lui (40,82 %).

Sens de *entendre*	*La Meuse*		*Le Robert*		Corpus général	
volonté	179	56,65 %	28	17,72 %	207	43,67 %
perception	129	40,82 %	129	81,65 %	258	54,43 %
écoute / attention	5	1,58 %	1	0,63 %	6	1,27 %
juger	1	0,32 %	0	0,00 %	1	0,21 %
comprendre	2	0,63 %	0	0,00 %	2	0,42 %
total	316	100,00 %	158	100,00 %	474	100,00 %

Tableau 4. Les sens du verbe entendre *avec le patron syntaxique [*V* [V]]*

De manière générale, la distinction entre les sens associés au patron [*V* [V]] relève de différences à la fois syntaxiques et sémantiques. Parmi les constructions les plus fréquentes, on relève les constructions présentées ci-dessous.

Exemple 16. [*V* [VCompl]] ie » ? France M. (Vaux-sur-Chèvremont), au chômage, est divorcée. « [J'ai entendu dire que je pourrais bénéficier d'une partie de la pension de mon ex-mari] (entendre_LaMeuse_121001_10)

Dans des cas de ce genre, la catégorie sémantique du verbe de la première complétive fait souvent, à elle seule, basculer le sens de la phrase. Les verbes de parole (*dire, répondre*…) ou les verbes qui impliquent un bruit (*ronronner, remuer, pleurer*…) donnent à *entendre* le sens de la perception auditive et sont d'ailleurs très courants : 80,95 % des cas de [*V* [V Complétive]].

Les mêmes remarques sur le type de verbe présent dans la complétive valent pour les constructions [*V* [V SN]]. Par contre, cette fois, les deux corpus présentent des comportements opposés quant au sens qui est le plus souvent associé à ce genre de construction. L'analyse du registre neutre indique une préférence très marquée pour le sens de la volonté : 91,74 % des constructions [*V* [V SN]] de ce registre signifient *vouloir*. Inversement, dans le corpus littéraire, 80 % de ces constructions signifient *percevoir par l'audition*.

Exemple 17. [*V* [V SN]] Toute la Vallée de la Meuse sera desservie [Proximus entend étendre la couverture sur la zone frontalière.] Inutile, pour les déte (entendre_LaMeuse_201201_07)

Pour la construction [*V* [V SP]], il s'agit à nouveau, comme dans les constructions des exemples précédents, de catégoriser le verbe de la complétive pour déterminer le sens à donner au verbe *entendre*. Le corpus littéraire utilise davantage cette construction avec le sens de la perception auditive (82,14 % avec cette construction dans ce corpus), tandis que le corpus neutre fait un usage partagé de la construction (perception auditive, 47,32 % et volonté, 52,68 %).

Exemple 18. [*V* [V SP]] é. N'existait aucun problème entre les joueurs et l'entraîneur. [Le groupe entendait profiter de la visite de Beveren parce que cette formation se situait] (entendre_LaMeuse_181001_03)

7.3. *Le patron [AuxPass *V*]*

Cette construction du verbe *entendre* apparaît au troisième rang des constructions les plus employées dans le corpus avec 217 cas, c'est-à-dire 10,43 %. Les deux corpus offrent un comportement identique quant à son utilisation dans le sens de *accepter* (environ 20 %). Pour le reste des sens, il s'agit de la construction pour laquelle les plus grandes disparités existent entre les deux types de corpus. Effectivement, le sens le plus souvent associé à l'emploi passif du verbe est l'*écoute officielle*. Or, il faut savoir que seul le corpus *La Meuse* comporte des

phrases de type [AuxPass *V*] ayant ce sens (40,66 % de ce corpus). Le sens que le corpus littéraire associe le plus souvent à cette construction est *la relation - accord* (40 %) alors que le corpus journalistique l'emploie avec nettement moins de fréquence (8,79 %).

Sens de *entendre*	***La Meuse***		***Le Robert***		**Corpus général**	
écoute officielle	74	40,66 %	0	0,00 %	74	34,10 %
écoute - attention	49	26,92 %	1	2,86 %	50	23,04 %
accepter	35	19,23 %	7	20,00 %	42	19,35 %
relation - accord	16	8,79 %	14	40,00 %	30	13,82 %
perception auditive	6	3,30 %	4	11,43 %	10	4,61 %
comprendre	1	0,55 %	9	25,71 %	10	4,61 %
écoute favorable	1	0,55 %	0	0,00 %	1	0,46 %
total	182	100,00 %	35	100,00 %	217	100,00 %

Tableau 5. Les sens du verbe entendre *avec le patron syntaxique [AuxPass *V*]*

Les citations de *La Meuse* montrent en fait trois exploitations sémantiques de la construction passive : outre *l'écoute officielle*, on relève *l'écoute – attention* (26,92 %) et *accepter* (19,23 %). Les trois plus fortes exploitations de cette même construction dans le corpus littéraire sont la *relation – accord*, *comprendre* et *accepter*. Les disparités entre les deux corpus permettent de conclure qu'elles sont de l'ordre du registre de langue.

Pour l'utilisation de la construction [AuxPass *V*], nous aborderons certains cas pour lesquels on relève des indices syntaxiques associés à des catégories sémantiques. Cette association d'indices permet de délimiter clairement le sens attribué à la construction. Ainsi, dans le corpus *La Meuse*, la présence d'un syntagme prépositionnel en *par* (21,43 % de ce corpus dans les construction de type [AuxPass *V*]) détermine dans 38 phrases sur 39 le sens de l'écoute. Cette écoute peut être caractérisée d'officielle lorsque le nom introduit par la préposition indique un représentant de la justice (*juge d'instruction, policier*…).

Exemple 19. se sont enfuis à pied vers les garages. [L'amie de M. Leroy a évidemment été entendue par les enquêteurs] mais n'a pu leur donner aucune indication précise. (entendre_LaMeuse_041201_07)

Le seul cas où la construction [AuxPass *V* SPpar] signifie la perception auditive, c'est lorsque le nom introduit par la préposition est un moyen de communication :

Exemple 20. amie de la victime, assisteront oralement à la scène. [«La conversation a été entendue par le téléphone] et il ne semble pas que la fille était consentante pou (entendre_LaMeuse_201001_27)

Dans le corpus littéraire, la préposition utilisée dans cette construction est le *de* et le sens qui y est associé est la compréhension. Outre la présence d'un syntagme prépositionnel à la suite du verbe, on trouve une autre construction bien représentée dans les deux corpus. Il s'agit de l'auxiliaire passif et du verbe suivi d'une complétive. Le corpus neutre fait un usage courant de l'auxiliaire au participe présent (86,84 % de ces constructions [AuxPass *V* Compl]) et lui attribue le sens de *accepter*, un usage beaucoup plus restreint dans le corpus littéraire (2 cas sur 13).

Exemple 21. at a mis en place une structure limpide. Le club en avait bien besoin, [étant entendu que ce n'était pas tant la bonne volonté qu'une solide organisation qu] (entendre_LaMeuse_101001_07)

7.4. *Le patron [CLse *V*]*

L'utilisation du verbe *entendre* en construction pronominale peut être associée à 5 sens. Les deux corpus présentent des comportements différenciés par rapport à cette construction. Seul le sens de la perception auditive est employé avec la même fréquence (27,56 % dans *La Meuse* et 29,09 % dans *Le Robert*). L'expression de la relation réciproque est la plus employée dans le corpus plus neutre (45,67 %) alors qu'elle n'est que de 16,36 % dans le corpus littéraire.

Sens de *entendre*	***La Meuse***		***Le Robert***		**Corpus général**	
relation réciproque	58	45,67 %	9	16,36 %	67	36,81 %
perception auditive	35	27,56 %	16	29,09 %	51	28,02 %
relation - accord	19	14,96 %	16	29,09 %	35	19,23 %
comprendre	6	4,72 %	12	21,82 %	18	9,89 %
habileté	9	7,09 %	2	3,64 %	11	6,04 %
total	127	100,00 %	55	100,00 %	182	100,00 %

Tableau 6. Les sens du verbe entendre *avec le patron syntaxique [CLse *V*]*

Somme toute, les emplois littéraires présentent une polysémie entre 4 sens répartis en deux blocs d'usage (la perception auditive et la relation – accord d'une part, la compréhension et la relation réciproque d'autre part) tandis que le corpus journalistique présente une polysémie davantage polarisée. La relation réciproque est le premier choix, viennent ensuite la perception auditive et la relation – accord et enfin, moins employés, l'habileté et la compréhension.

Nous signalons ici deux différences notables entre les deux corpus. Tout d'abord, l'expression de l'habileté au moyen de ce patron syntaxique est très peu exploitée par le corpus littéraire. À l'inverse, le corpus neutre en fait davantage usage. La construction que l'on trouve dans 8 cas sur 9 consiste à insérer un clitique indirect accompagné soit d'un syntagme prépositionnel soit d'une comparative : *il s'y entend en matière de…, il s'y entend comme personne.* La seconde différence à noter est l'expression de la relation réciproque. Elle se fait souvent au moyen d'un adverbe de qualité (le plus souvent l'adverbe *bien*) dans le corpus *La Meuse* : 51,72 % des cas de *relation réciproque* (30 cas sur 58). Tandis que, dans le corpus littéraire, cette même construction ne représente qu'un cas sur 9. Essentiellement, dans ce registre de langue, ce sens se construit avec un adverbe de qualité et en plus avec un syntagme prépositionnel : [CLse *V* SPavec].

7.5. *Le patron [*V* [x V]]*

Cette construction de *entendre* est majoritairement associée au sens de la perception auditive, et cela autant dans le corpus général que dans chacun des deux corpus. Les citations littéraires montrent malgré tout une préférence très marquée pour ce sens (85 %). Le deuxième sens qui se trouve être associé à cette construction est celui de la volonté. Il faut noter que les emplois littéraires font un moindre usage de la construction [*V* [x V]] dans le sens de vouloir (15 %) que ne le fait le corpus journalistique (25 %). Le troisième sens que prend le verbe *entendre* dans cette construction est *l'écoute.* Il n'est utilisé de cette manière que dans les phrases tirées de *La Meuse* (8,82 %). Par ailleurs, cet emploi se rattache au domaine de la perception sensorielle, qui est donc le domaine sémantique le plus associé à cette construction.

Sens de *entendre*	*La Meuse*		*Le Robert*		Corpus général	
perception auditive	45	66,18 %	34	85,00 %	79	73,15 %
volonté	17	25,00 %	6	15,00 %	23	21,30 %
écoute - attention	6	8,82 %	0	0,00 %	6	5,56 %
total	68	100,00 %	40	100,00 %	108	100,00 %

Tableau 7. Les sens du verbe entendre *avec le patron syntaxique [*V* [x V]]*

Le sens de la perception auditive est associé aux constructions plus spécifiques [*V* [SN V]] et [*V* [Pronom V]]. Dans ces cas, le nom ou le pronom est le sujet du verbe de la complétive infinitive. Il représente des êtres humains (*un riverain, certains*…) et est accompagné d'un verbe d'expression orale tel que *dire, se plaindre, demander, crier*... Lorsque cette même construction signifie *l'écoute-attention*, c'est tout simplement que l'association entre le sujet et le verbe infinitif fait basculer la phrase dans un contexte de rituel d'écoute : par exemple, *entendre quelqu'un interpréter une composition*. Par contre, lorsque le pronom ne représente plus des êtres humains (*tout, y*…) ou que la construction spécifique est soit [*V* [clitique direct V]] soit [*V* [clitique indirect V]], le sens est alors celui de la volonté.

Exemple 22. [Mais, pour si précieux qu'il le tînt, Bonaparte entendait ne pas le mettre au pinacle.] Louis MADELIN, Talleyrand, II, XII. (entendre_GrandRobert_cit20_720)

7.6. *Le patron [*V* Compl]*

Cette classe de phrases présente essentiellement deux significations : la compréhension (61,17 %) et la perception auditive (33,01 %). Le corpus littéraire présente une forte préférence pour l'association entre le sens de *comprendre* et cette construction (80 %). Dans le corpus journalistique, le sens de *comprendre* reste le plus associé à cette construction (53,42 %), mais le sens de la perception auditive est, lui aussi, très employé (39,73 %).

Sens de *entendre*	*La Meuse*		*Le Robert*		Corpus général	
comprendre	39	53,42 %	24	80,00 %	63	61,17 %
perception auditive	29	39,73 %	5	16,67 %	34	33,01 %
volonté	4	5,48 %	1	3,33 %	5	4,85 %
accepter	1	1,37 %	0	0,00 %	1	0,97 %
total	73	100,00 %	30	100,00 %	103	100,00 %

*Tableau 8. Les sens du verbe « entendre » avec le patron syntaxique [*V* Compl]*

L'expression de la compréhension dans le registre journalistique se fait de deux manières : au moyen de l'auxiliaire modal *faire* ou *laisser* (82,05 % des cas) ou en ajoutant un syntagme prépositionnel en *par* (17,95 % des cas). Dans le registre littéraire, le sens de compréhension est également associé à la construction avec un auxiliaire modal (45,83 % des cas), tandis l'utilisation de l'adverbe *bien* portant sur le verbe *entendre* apparaît comme le deuxième choix pour donner au verbe *entendre* le sens de *comprendre.*

8. Conclusion

Nous avons constaté que le verbe *entendre* se construit au moyen de 9 patrons syntaxiques, ayant chacun des fréquences d'emploi variées. Le patron le plus fréquent est, de loin, la construction transitive directe, ensuite la construction avec une complétive infinitive [*V* [V]]. Ces deux constructions totalisent à elles seules presque 70 % des emplois du corpus. L'intersection de contraintes syntaxiques plus spécifiques et de préférences sélectionnelles

pour les arguments du verbe (notamment les types de substantifs) a permis de tracer les lignes directrices qui gèrent l'association de significations à des constructions particulières. Ainsi, dans les différents patrons analysés, l'association du verbe *entendre* avec un syntagme nominal ou un verbe de type auditif confère généralement au verbe *entendre* un sens relevant du domaine de l'audition. Des contraintes syntaxiques en sus du patron syntaxique peuvent aider à mieux guider le choix du sens à donner au verbe *entendre* dans un contexte. Nous avons par exemple constaté que le patron [*V* Compl], lorsqu'il était utilisé avec des auxiliaires modaux tels que *laisser* ou *faire*, prenait le sens de *comprendre*.

Par ailleurs, l'étude d'un corpus comprenant deux registres de langue a permis d'établir des différences de comportement dans les choix opérés parmi les emplois du verbe *entendre.* Dans la majorité des cas, les emplois diffèrent plutôt par une polarisation plus marquée dans le registre littéraire. En effet, ce registre montre des préférences plus nettes pour associer un patron syntaxique à un sens du verbe *entendre*, tout en utilisant le même éventail de sens que le registre neutre. Au contraire, le registre neutre présente davantage de répartition dans les fréquences d'association entre les différents sens du verbe *entendre* et chaque patron syntaxique utilisé. Une seule exception : la construction [CLse *V*] est davantage polarisée (envers le sens de la relation réciproque) dans ce registre que dans le registre littéraire.

Références

Clark S. et Weir D. (2002). Class-Based Probability Estimation using a Semantic Hierarchy. *Computational Linguistics*, vol. (28/2) : 187-206.

Levin B. (1993). *English Verb Classes and alternations. A Prelimanary Investigation*. The University of Chicago Press.

Mann V. et Liberman P. (1983). Some differences between phonetic and auditory modes of perception. *Cognition,* vol. (14) : 211-235.

Manning C.D. et Schütze H. (2000). *Foundations of Statistical Natural Language Processing*. MIT Press.

Perez I. (1994). Les agnosies auditives : une analyse fonctionnelle. In McAdams S. et Bigand E. (Eds), *Penser les sons : psychologie cognitive de l'audition*. PUF : 215-248.

Piron S. (2002). Les verbes de perception génériques en français. Une approche historique du lexique. Communication donnée à l'ACFAS, *Discipline Linguistique*, 14-15 mai 2002, Université Laval, Québec, Canada.

Resnik P. (1997) Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?*

Pré-analyse de corpus

Thierry Poibeau

Laboratoire d'Informatique de Paris Nord (LIPN)
Université Paris 13 – Av. J-B. Clément – 93430 Villetaneuse – France
poibeau@lipn.univ-paris13.fr

Abstract

Most Natural Language Processing tools need homogeneous corpora in order to deliver relevant results. However, such corpora are rarely available in industrial and applicative contexts. This paper presents an original approach for preparing corpora in order to obtain useful amount of texts. The presented techniques are based on statistical and surface linguistic analysis. We present these techniques and an experiment in the information extraction domain. We demonstrate the different techniques and evaluate their interest for the task.

Résumé

Les applications de Traitement Automatique des Langues nécessitent le plus souvent des corpus homogènes pour fournir des résultats pertinents. De tels corpus sont rarement disponibles dans des contextes applicatifs ou industriels. Cet article propose une approche originale pour préparer les corpus et obtenir des masses de textes utilisables. Les techniques présentées vont des statistiques à l'analyse linguistique de surface. Les techniques sont présentées puis appliquées au domaine de l'extraction d'information. Nous montrons l'intérêt de ces techniques et nous en donnons une évaluation fondée sur leur pertinence pour la tâche visée.

Mots-clés : pré-analyse de corpus, analyse linguistique de surface, approches statistiques

1. Introduction : analyse et représentation du contenu des textes

Le Traitement Automatique des Langues s'est beaucoup intéressé dans les décennies passées à la compréhension de textes. La compréhension, au moins partielle, est nécessaire pour le filtrage, l'aide à la décision et la traduction. Devant la difficulté de la tâche, une approche fondée sur une analyse partielle a été proposée, à travers notamment les techniques d'extraction d'information. Il ne s'agit plus de comprendre tout le texte mais d'extraire un certain nombre d'informations pertinentes par rapport à un formulaire prédéfini en fonction de la tâche visée.

Même si l'extraction est actuellement une technologie bien maîtrisée, la question de la mise au point des systèmes reste un problème crucial. Plusieurs auteurs ont proposé le recours à l'apprentissage pour l'acquisition semi-automatique de lexiques et de grammaires d'extraction mais ces techniques laissent de côté plusieurs aspects importants : comment savoir si un ensemble de textes se prête bien à une analyse en vue d'une application d'extraction d'information ? L'ensemble de textes considéré porte-t-il sur des faits avérés ou s'agit-il de déclarations ou de rumeurs ? Quelle est la part de modalité ou de négation, phénomènes que l'on sait mal analyser, surtout sur du texte « tout venant » ? Ces phénomènes posent des problèmes majeurs aux stratégies d'analyse locale telles qu'elles sont mises en œuvre dans le cadre des applications d'extraction.

Dans de nombreux cas, les concepteurs d'applications partent de corpus hétérogènes faits de plusieurs types de textes traitant de différents domaines. Or, on sait que l'extraction nécessite des corpus homogènes et qu'il faut au préalable élaborer un ensemble de textes pertinents pour pouvoir l'analyser puis en extraire des connaissances. Il existe heureusement des outils permettant de procéder à une pré-analyse de corpus afin d'en déterminer les principales caractéristiques. Les techniques de cartographies permettent notamment d'opposer plusieurs sous-corpus d'après certains traits caractéristiques. Cet article montre une utilisation d'outils d'analyse de corpus – statistiques et linguistiques – dans cette perspective applicative.

Nous présentons tout d'abord les techniques à l'œuvre. Nous détaillons ensuite les différentes expériences effectuées et nous montrons ainsi la complémentarité entre les approches et les outils utilisés pour cette tâche. Nous procédons ensuite à une évaluation des différentes techniques mises en œuvre.

2. Techniques de pré-analyse de corpus

Les techniques mises en œuvre pour la pré-analyse de texte sont surfaciques, pour des raisons évidentes d'efficacité. On trouve essentiellement des méthodes statistiques mais aussi quelques analyses de type linguistique.

2.1. *Méthodes quantitatives et analyse de spécificités*

Les méthodes statistiques permettent des traitements rapides sur de gros corpus. Ce type de méthode peut ainsi être valablement utilisé pour la partition d'un corpus pour peu que l'on dispose d'une base d'entraînement. Dans le cadre qui nous occupe (pré-analyse de textes dans la perspective d'applications d'extraction d'information) cette condition est généralement remplie. Les analystes sont le plus souvent des experts ayant emmagasiné des textes pertinents dans leur domaine de spécialité.

L'analyse de spécificités peut porter sur les formes du texte, sur des formes lemmatisées ou sur une analyse de spécificités. Nous utilisons l'outil LEXICO d'A. Salem *et al.* (2002), qui permet de combiner analyse de spécificités et recherche de segments répétés (« suite de formes dont la fréquence est supérieure ou égale à deux dans le corpus », d'après (Lebart et Salem, 1994)). Ces formes sont souvent plus représentatives d'un corpus que les formes simples. On peut alors procéder à un calcul de spécificités positives et négatives : « pour un seuil de spécificité fixé, une forme *i* et une partie *j* données, la forme *i* est dite spécifique positive (resp. négative) de la partie *j* (ou forme caractéristique de cette partie) (...) si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures (resp. inférieures) à la sous-fréquence constatée est inférieure au seuil fixé au départ. » (repris de (Lebart et Salem, 1994)).

Une analyse factorielle de correspondance (AFC) permet de projeter les résultats sur un plan et de faire apparaître certains regroupements. Si ce type d'étude donne souvent des résultats difficiles à interpréter, nous montrons dans ce qui suit que l'AFC peut permettre de dégager automatiquement des regroupements de textes intéressants au sein d'un corpus hétérogène. Nous utilisons cette technique de manière très simple pour faire apparaître des sous-corpus au sein d'un ensemble hétérogène de textes.

2.2. *Analyse linguistique de surface*

L'analyse linguistique vise à préciser et à affiner les résultats de l'analyse statistique. Nous avons vu dans l'introduction que tous les textes ne se prêtent pas bien à l'extraction d'information. Si le texte comporte trop de modalités, si l'analyse linguistique risque d'être

troublée par des phénomènes de portée (de quantifieurs, de négations, etc.), il importe de se montrer prudent sur la valeur du résultat obtenu. On sait en effet qu'il s'agit là de marqueurs de prise de distance par rapport au fait, pouvant introduire un doute ou nuancer l'information rapportée. L'analyse de spécificités de textes non événementiels (textes qui ne relatent pas que des faits mais qui incluent aussi des commentaires et des prises de position) met souvent en avant des formes modales et des négations qui doivent être prises en compte pour juger de la pertinence globale que l'on cherche à atteindre. Si l'analyse de spécificités montre trop de formes modales, alors l'analyse doit être regardée avec prudence et un retour au texte peut s'avérer nécessaire.

Enfin, une fois que l'on a obtenu un corpus homogène et pertinent, il peut être utile de filtrer les sous-parties du corpus pertinentes pour la tâche visées. L'expérience acquise dans le cadre des conférences américaines d'évaluation MUC a montré que, souvent, moins de 10 % du corpus était utile (MUC, 1995). Dans certains cas extrêmes comme la recherche de relations d'interaction au sein d'un corpus de génomique, environ 3 % du texte est pertinent. Des techniques simples et surfaciques peuvent permettre d'effectuer ce filtrage (Bessières *et al.*, 2001). Nous avons précédemment expérimenté une technique fondée sur la densité sémantique dans (Poibeau, 2003) : un filtrage thématique peut aisément être implanté au moyen de mots clés (généralement fournis par l'utilisateur) enrichis sémantiquement par le recours à un réseau de type Wordnet. Nous montrons ici comment l'analyse de formes particulières comme les entités nommées peut également favoriser ce filtrage.

3. Expériences

Nous avons constitué un corpus composé pour une moitié de dépêches AFP et pour une autre moitié de courriers électroniques. Afin d'envisager un traitement automatique, il est nécessaire de pouvoir partitionner cet ensemble de textes en deux sous-ensembles distincts. Pour ce faire, nous utilisons le logiciel LEXICO décrit dans la partie 2.1.

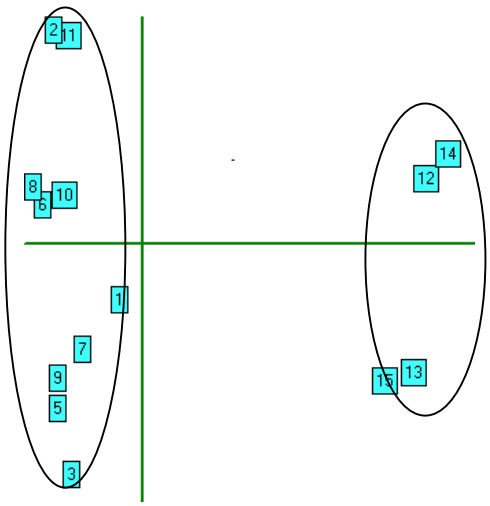


Figure 1. Caractérisation de corpus par analyse statistique.
On voit clairement apparaître deux partitions, entourées d'un trait noir sur le graphique.

Nous donnons ci-dessus une carte produite par analyse factorielle de correspondance produite suite à une analyse de spécificités du corpus. Le résultat fait clairement apparaître deux ensemble de points, répartis de part et d'autre d'un axe vertical. Un retour au texte permet de voir que l'ensemble de gauche correspond aux courriers électroniques et que l'ensemble de droite correspond aux dépêches AFP. L'analyse ne prend pas tant en compte des oppositions d'ordre lexical que des oppositions d'ordre morpho-syntaxique, voire typographique. Il est

aussi possible d'opposer les textes d'un même domaine mais appartenant à des genres différents (cf. Biber, 1995 et 1998).

L'analyse de la carte obtenue est révélatrice du contenu des documents et de la façon dont ils se prêtent à l'extraction d'information. L'ensemble de points situé à gauche (les courriers électroniques) est marqué par un ensemble de facteurs peu propices à l'extraction : présence de pronoms de la première personne du singulier, d'opérateurs de modalités, etc. Quand on analyse l'ensemble des dépêches AFP au regard de ce corpus, les spécificités négatives (formes sous-représentées dans le corpus AFP par rapport au corpus de courriers électroniques) sont à cet égard très éclairantes :

Terme	Fréquence	Spécificité
Je	103	-10
Windows	47	-9
ai	41	-9
J'	29	-9
Comment	53	-8
dois	23	-8
un	122	-7
puis	18	-6
mon	21	-6
insérer	14	-6

Figure 2. Analyse de spécificités. On voit nettement apparaître que les formes qui pourraient laisser planer un doute sur la faisabilité de l'extraction (présence de verbes à la première personne, de modaux, ...) apparaissent dans la série des formes sous-représentées. Il s'agit donc d'un corpus pertinent pour une application d'extraction.

Il est ainsi possible, en mettant en œuvre des moyens simples et automatiques, de caractériser grossièrement des corpus afin de déterminer leur adéquation à une tâche donnée. Il serait sans doute possible d'aller plus loin et de prédéfinir une série d'expressions ou de marqueurs se prêtant mal à une application d'extraction. Cette stratégie est proche de l'*exploration contextuelle* définie par ailleurs par Minel (2003). Le système ne peut bien évidemment pas prendre de décision automatique, le choix final revenant à l'utilisateur. C'est à lui de juger de la pertinence du texte, de voir si une application d'extraction est envisageable[1] mais une série de marqueurs pourrait quand même être définie *a priori*.

L'extraction d'information est largement fondée sur l'analyse de faits impliquant des entités nommées. De fait, les passages impliquant des entités nommées sont souvent pertinents pour les tâches d'extraction. Il existe aujourd'hui plusieurs systèmes permettant de repérer de manière relativement pertinente des entités nommées. Des techniques d'analyse de densité et de visualisation de données peuvent ensuite faire apparaître de manière claire les passages

[1] Cette tâche suppose un niveau d'expertise technique qui laisse supposer que des utilisateurs experts ou des administrateurs ayant à la fois des connaissances sur le domaine technique visé, et des connaissances en modélisation linguistique. Il semble encore illusoire de vouloir donner les outils de conceptions d'outils linguistiques directement aux utilisateurs finaux, sauf pour des tâches limitées.

marqués par une forte concentration d'entités. La figure ci-dessous montre une telle analyse sur une partie de notre corpus.

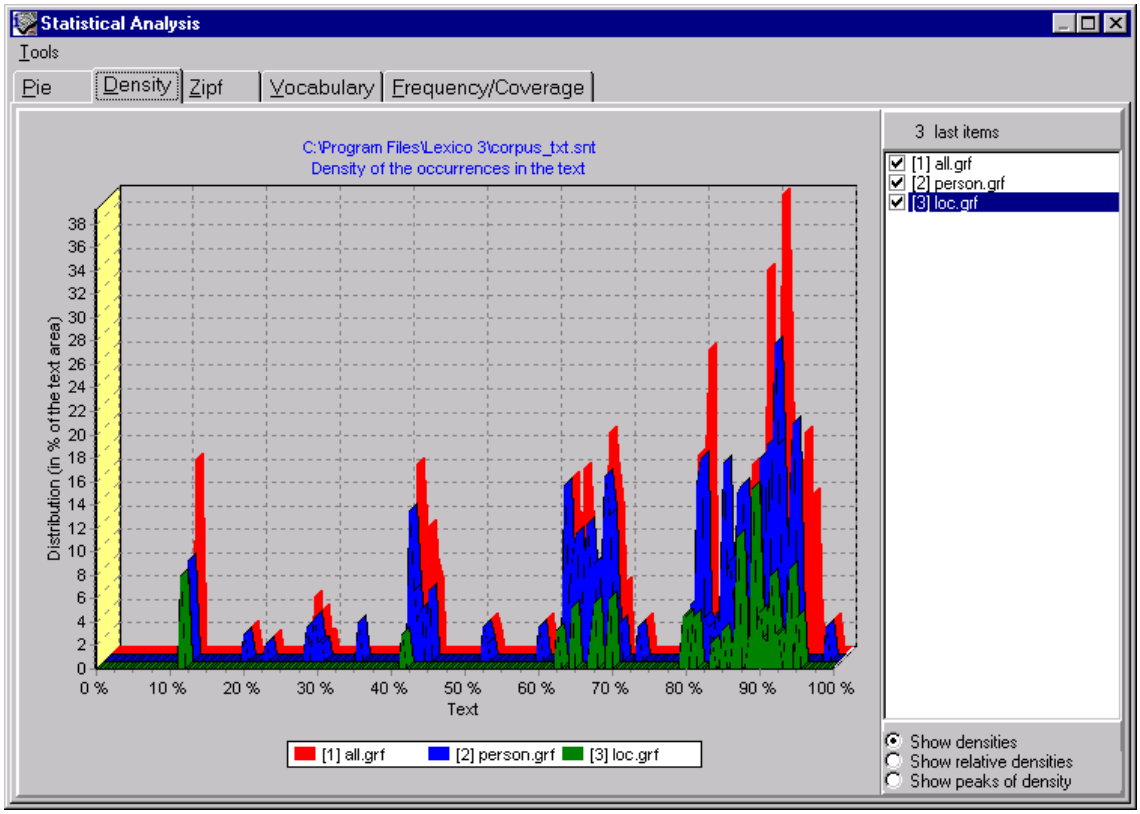


Figure 3. Analyse de densité portant sur les entités nommées du corpus. Une première étape a permis de marquer au sein du corpus les entités reconnues. L'outil statistique ici présenté (module statistique d'Intex, cf. Silberztein 1993) permet de repérer les zones marquées par une forte densité d'entités.

La figure 3 présente des courbes de densité d'entités nommées pour le corpus étudié. La courbe en supérieure rassemble le résultat de l'ensemble des entités, les deux courbes inférieures révèle pour l'une la présence de noms de personnes et pour l'autre la présence de noms de lieux. Les trois courbes sont largement homogènes sur tout le corpus.

4. Évaluation

La mise au point du système décrit dans la section précédente a porté sur un ensemble réduit de textes. Elle a été reproduite sur un corpus plus volumineux afin d'obtenir des chiffres plus fiables.

4.1. *Le corpus*

Un ensemble de dépêches provenant de la société FirstInvest a été utilisé comme corpus de référence. Il s'agit d'un « corpus de suivi » (Habert *et al.*, 1997) dans la mesure où l'on a affaire à un amas de textes qui grossit chaque jour, par opposition aux corpus généralement utilisés en linguistique, qui sont constitués une fois pour toutes sur un thème donné.

Afin d'établir un corpus de référence, un ensemble de 2000 dépêches réparties en 26 thèmes différents nous a été fourni. Nous avons extrait de ce corpus les dépêches se rapportant au thème des rachats d'entreprises (acquisitions, cessions, prises de participation, offres publiques d'achat, …). Le corpus obtenu à partir du thème « rachat d'entreprises » est composé de

303 dépêches (soit 56 000 mots). Ce corpus a été analysé manuellement pour pouvoir procéder à une évaluation de type MUC mais ici nous nous contentons du corpus brut, qui est suffisant pour nos expériences. Ce corpus est divisé en deux : un ensemble de 243 textes a constitué le corpus d'entraînement (45 334 mots), les 60 textes restants formant le corpus de test (10 666 mots).

4.2. *Évaluation*

Le corpus correspondant au thème « rachat d'entreprise » a été mêlé à un ensemble de textes de taille correspondante sur le thème « rumeur » (environ 15 000 mots) puis sur le thème « nouvelle économie » (environ 24 000 mots). Les résultats sont contrastés. Nous avons procédé à deux expériences, en mêlant tout d'abord les textes du thème « rachat d'entreprises » à ceux du thème « rumeur » puis au thème « nouvelle économie ». Nous partitionnons ensuite le corpus en deux en utilisant Lexico comme précisé ci-dessus. Il est alors possible de mesurer le nombre de textes bien catégorisés (c'est-à-dire, correspondant à la classe initiale). Nous obtenons les résultats suivants :

	Bien classés	Mal classés
Thème « rachat d'entreprise »	91 %	9 %
Thème « rumeur »	79 %	21 %

	Bien classés	Mal classés
Thème « rachat d'entreprise »	82 %	18 %
Thème « nouvelle économie »	77 %	23 %

On voit sur ces résultats que les textes sont globalement bien classés, même si des erreurs subsistent. Le premier constat est que le système de classement ainsi constitué fonctionne mieux sur les thèmes « rachat d'entreprise » *versus* « Rumeur » que sur « rachat d'entreprise » *versus* « Nouvelle économie ». Comme on peut aisément l'imaginer, les textes du thème « Rumeur » comportent en effet des modaux et des adverbes de prises de distance, qui sont à l'inverse peu présent dans le thème « rachat d'entreprise ». Les termes employés dans le corpus « nouvelle économie » sont en revanche moins discriminants. On obtient donc des résultats moins bons : 23 % des messages se rapportant à ce corpus étant classés dans le thème « rachat d'entreprise » dans notre expérience. Il faut cependant relativiser ces résultats : un texte peut appartenir à un thème donné et être aussi pertinent pour un autre thème.

Nous avons ensuite testé la sélection de passages à partir des entités nommées. Un filtrage uniquement conçu sur les entités nommées n'est pas performant si l'on compare les résultats obtenus à partir de cette méthode avec les résultats d'une méthode fondée sur un ensemble de mots clés. Des passages pertinents ayant une faible concentration d'entités nommées ne sont pas repérés et, à l'inverse, les entités ne semblent pas être un critère suffisant pour retenir un passage. Une étude manuelle des résultats montre cependant que les passages comportant des entités nommées sont plus pertinents que les passages sans entités nommées. Il est donc possible d'éliminer automatiquement un certain nombre de passages non pertinents. La proportion de texte retenu peut ainsi passer de 20 % à 15 % du texte original sans conséquence majeure pour la tâche suivante (la phase d'extraction proprement dite). On a ainsi un gain de temps de traitement du fait de l'amélioration du filtrage initial marqué par un nombre infé-

rieur de textes à traiter (passage de 20 à 15 % de la masse textuelle initiale, soit une diminution de 25 %).

5. Conclusion

Nous avons présenté dans cet article quelques expériences utilisant des méthodes rapides afin de caractériser *a priori* un corpus dans la perspective d'une application d'extraction d'information. Nous avons ainsi montré qu'il était aisé d'obtenir une analyse statistique (à partir d'outils comme Lexico) permettant de partitionner un corpus afin d'obtenir un ensemble de textes homogène. De plus certains indices linguistiques (modaux, négation, etc.) permettent de prédire de façon sommaire mais efficace si les textes en question peuvent se prêter à une application d'extraction. Dans certaines conditions, la pré-analyse des textes peut inciter à être prudent sur les résultats et à les valider manuellement. Nous avons enfin présenté un filtrage fin des données linguistiques en fonction de l'analyse des entités nommées. Les résultats que nous avons présentés offrent une analyse fine des phénomènes en jeu. Ce travail reste à compléter afin d'obtenir des systèmes d'extraction ou de compréhension opérationnels sur des données réelles.

Références

Bessières P., Nazarenko A. et Nédellec C. (2001). Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques. In *4e Colloque International sur le Document Électronique* (CIDE'201) : 165-183.

Biber D. (1998). *Variation across speech and writing*. Cambridge University Press.

Biber D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press.

Habert B., Nazarenko A. et Salem A. (1997). *Les linguistiques de corpus*. Armand Colin.

Lafon P. (1984). Pour une présentation de la méthode et du calcul de la probabilité. In *Dépouillements et statistiques en lexicométrie*. Champion-Slatkine.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Minel J.-L. (2003). *Filtrage d'information*. Hermès.

MUC (1995). *Proceedings of the 6th Message Understanding Conference (MUC 6)*. Morgan Kaufmann.

Poibeau T. (2003). *Extraction d'information, du texte brut au web sémantique*. Hermès.

A. Salem *et al.* (2002). *Manuel d'utilisation de Lexico.* Université Paris 3.

Silberztein M. (1993). *Dictionnaires électroniques et analyse automatique de textes, le système INTEX.* Masson.

Des variables tensives inscrites dans le texte : une interprétation dynamique de l'A.F.C. dans l'analyse d'Alceste

Jean-Luc Pommier

Formateur à l'IUFM d'Auvergne, doctorant en Sciences de l'Éducation
Université Jean Monnet – Saint-Étienne
Centre de Recherche en Éducation – 35 rue du 11 novembre – 42023 Saint-Étienne cedex 2
Adresse personnelle : Chemin des fichardies, 63300 Thiers – jean-luc.pommier@wanadoo.fr

Abstract

Tension variables inscribes in the text : a dynamic interpretation of F.A.C. in the analysis of Alceste. In the use of the Alceste software, the factorial analysis of correspondences (F.A.C.) is used to link the resulting classes of the hierarchical downward classification. We will try to go beyond the common interpretation of this software by exploiting the results in a dynamic perspective. The lines are interpreted as "tension variables", (variables) which analysis enables to reveal, putting under tension opposed (or even contrary) values. Classes are not only the object of a Contents Analysis (which is their most common interpretation) but they become components of the discourse, linked together and directed in relation to the tension variables. These links are difficult to name because they are potentialities. Their different names give access to the density of a semiotic analysis. As a matter of fact, according to the setting of the analysis, the lines will be called : pragmatic line and cognitive line, space-time line and identity line... We will use our corpus : a selection of help requests from primary school-teachers for some of their students with special needs. These requests are sent to a group of school-professionals in special needs called RASED. The analysis of F.A.C. shows the space for a missing discourse intricate with the tensions of the two lines. The representation of clouds shows closenesses and remotenesses, thus revealing a "missing discourse" from time to time, a discursive and non-updated potentiality. What isn't told at school? Does school-handicap include a part of taboo? Alceste gives access to hypotheses, which will be tested by other analyses, semio-narrative analyses.

Keywords : Alceste, F.A.C., dynamic, tension variables, school-discourse.

Résumé

Dans l'utilisation du logiciel 'Alceste', l'analyse factorielle de correspondances (A.F.C.) est utilisée pour articuler les classes obtenues par la classification descendante hiérarchique. Nous allons essayer de dépasser l'interprétation courante de ce logiciel en exploitant les résultats dans une perspective dynamique. Les axes sont interprétés comme des « variables tensives », des variables que l'analyse permet de révéler, qui mettent en tension des valeurs opposées voire contraires. Les classes ne sont plus seulement l'objet d'une Analyse de Contenu (leur interprétation la plus courante) mais deviennent des composantes du discours articulées entre elles et orientées par rapport aux variables tensives. Les axes sont difficiles à nommer parce que ce sont des potentialités. Leurs différentes nominations donnent accès à l'épaisseur d'une analyse sémiotique. En effet, selon le cadre de l'analyse, les axes seront nommés : axe pragmatique et axe cognitif, axe spatio-temporel et axe des identités … Nous utiliserons notre corpus : un recueil de demandes d'aide rédigées par des instituteurs pour certains de leurs élèves en difficulté[1]. Ces écrits sont adressés à une instance de professionnels de l'aide à l'école : le RASED. L'analyse de l'A.F.C. nous montre l'emplacement d'un discours absent intriqué dans les tensions des deux axes. La

[1] Thèse de doctorat en Sciences de l'Éducation en cours, Université Jean Monnet, Saint-Étienne, Directeur Marc Derycke, titre « La difficulté d'enseigner mise en scène dans les demandes d'aide (étude de fiches de signalement adressées au RASED) ».

représentation des nuages montre des rapprochements et des éloignements, révélant ainsi parfois un « discours absent », une potentialité discursive non actualisée. Qu'est-ce qui n'est pas dit à l'école ? La difficulté scolaire comporte-t-elle une part de tabou ? Alceste nous donne accès à des hypothèses qui seront testées par d'autres analyses, en l'occurrence, des analyses sémio-narratives.

Mots-clés : Alceste, A.F.C., dynamique, variables tensives, discours scolaire

1. Introduction

Nous avons tenté de pousser le plus loin possible l'interprétation de l'analyse d'Alceste. L'analyse factorielle de correspondances (A.F.C.) propose une présentation spatiale des contenus qui est orientée par des axes. Ces axes seront interprétés comme des variables tensives parce qu'ils orientent les valeurs dans une dialectique entre deux pôles en tension.

Cette conception nous donnera des éléments de discussion pour interpréter un espace vide entre deux *topoï*. Que peut signifier ce mouvement ? Si nous situons nos réflexions dans une perspective structuraliste, les analyses sur les unités linguistiques seront différentielles et oppositives. L'écart entre les classes sera lu comme un mouvement dans les tensions, un « discours absent » un hiatus entre les variables. Des traces de ce discours sont visibles : « travail » « comprendre » ou « compréhension ». Une lecture approfondie du rapport d'analyse nous aidera à faire apparaître la *difficulté d'enseigner*.

2. Le matériel utilisé : l'analyse d'Alceste

Le logiciel Alceste produit deux analyses successives, une classification descendante hiérarchique puis une analyse factorielle de correspondance. Nous allons utiliser une partie de notre travail de thèse de doctorat[2] pour l'exploiter d'une manière méthodologique. Ce travail est une Analyse de Discours portant sur des archives institutionnelles, des fiches de demande d'aide pour des élèves en difficulté à l'école primaire, fiches adressées par des instituteurs à l'instance d'aide de l'Éducation Nationale, le R.A.S.E.D.[3]. Nous avons utilisé Alceste de différentes manières et à différentes étapes : d'abord pour interroger la constitution du corpus global, puis pour valider la représentativité des échantillons (les cas choisis) et ensuite pour faire émerger des hypothèses de travail. Le corpus est constitué de 226 fiches rédigées entre 1997 et 2001 dans un secteur géographique de l'ordre d'une circonscription. Les textes sont obtenus par des réponses au questionnaire suivant :

1. Qu'est-ce qui vous incite à signaler l'enfant ?
2. Quel(s) problème(s) vous pose-t-il ?

Des aides ont-elles déjà été apportées ? 3. Dans la classe 4. Dans l'école 5. À l'extérieur

6. Que pouvez-vous dire de positif sur cet enfant ?
7. La famille est-elle informée des difficultés de l'enfant ? oui - non. Si oui, quelle est sa position ?
8. Quel type d'aide vous paraît souhaitable ? Qu'en attendez-vous ?

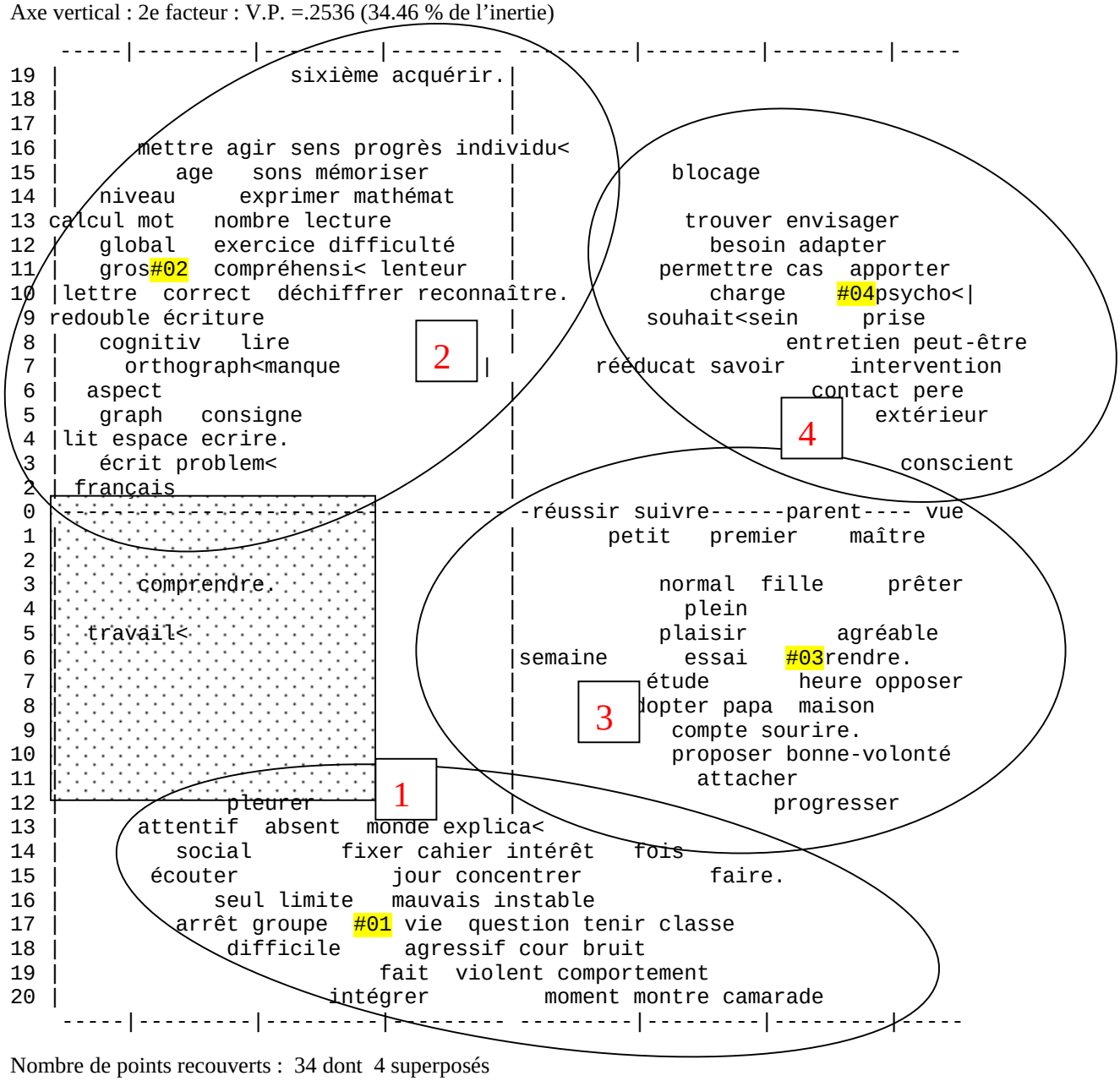
[2] Thèse de doctorat en Sciences de l'Éducation, Université Jean Monnet, Saint-Étienne, Directeur Marc Derycke, titre « La difficulté d'enseigner mise en scène dans les demandes d'aide (étude de fiches de signalement adressées au RASED) ».

[3] Réseau d'Aides Spécialisées aux Élèves en Difficulté. Ce groupe de travail est composé d'enseignants spécialisés dans l'aide psychologique, pédagogique.

Figure 1. Projection des mots analysés sur le plan 1 2 (corrélations)

Axe horizontal : 1e facteur : V.P. =.3196 (43.42 % de l'inertie)

Axe vertical : 2e facteur : V.P. =.2536 (34.46 % de l'inertie)



Nombre de points recouverts : 34 dont 4 superposés

Classe 1		Classe 2		Classe 3		Classe 4	
x	y nom	x	y nom	x	y nom	x	y nom
-16	-15 agiter	-26	13 oral	29	0 orthophon	23	10 orientation
-15	-15 période	-26	13 faible	30	0 famille	27	10 possible
-14	-15 repondre	-26	13 lent	32	0 maman	28	10 aide
-12	-15 perturb<	-26	13 échec	26	-1 voir	27	9 bilan
-11	-15 journée	-24	14 vraiment	26	-3 accepter	31	7 psycholog
-11	-16 agitation	-23	14 retard	29	-3 aimer		
-10	-16 refuser	-22	14 lire	25	-5 gentil		
-12	-17 remarquer	-25	12 phrase				
0	-20 parler	-24	12 math				
2	-20 chercher	-23	12 vocabulaire				
		-25	11 texte				
		-24	11 énorme				

2.1. *Analyse de Contenu, la Classification Descendante hiérarchique*

Les 4 classes obtenues sont nommées ainsi :

- Classe 1 : L'évaluation comportementale
- Classe 2 : L'évaluation cognitive
- Classe 3 : Demande à la famille
- Classe 4 : Demande aux professionnels

Ces classes sont des mondes lexicaux ou *topoï* organisant le contenu. Chaque mot est lié à une ou plusieurs classes selon un degré d'appartenance, le Khi2.

Classe 1 :	534	36. 109.	33.03	**3.66**	Y travail<
Classe 2 :	534	40. 109.	36.70	**8.39**	Y travail<

Par exemple, le lemme « travail » est très représentatif de la classe 2 (l'évaluation cognitive) avec un khi2 de 8.39 mais il est aussi lié à la classe 1, mais à un degré moindre, avec un khi2 de 3.66.

2.2. *Projection des mots sur le plan : l'analyse factorielle de correspondances*

Quatre nuages matérialisent les classes. Les signes # situent les centres de ces classes. Le rectangle nous montre un vide. C'est ce tableau qui sera questionné pour faire signifier les deux axes. Les axes seront interprétés comme des variables tensives faisant signifier l'espace donc à lé fois les classes dans leurs positions relatives et les espaces entre les classes.

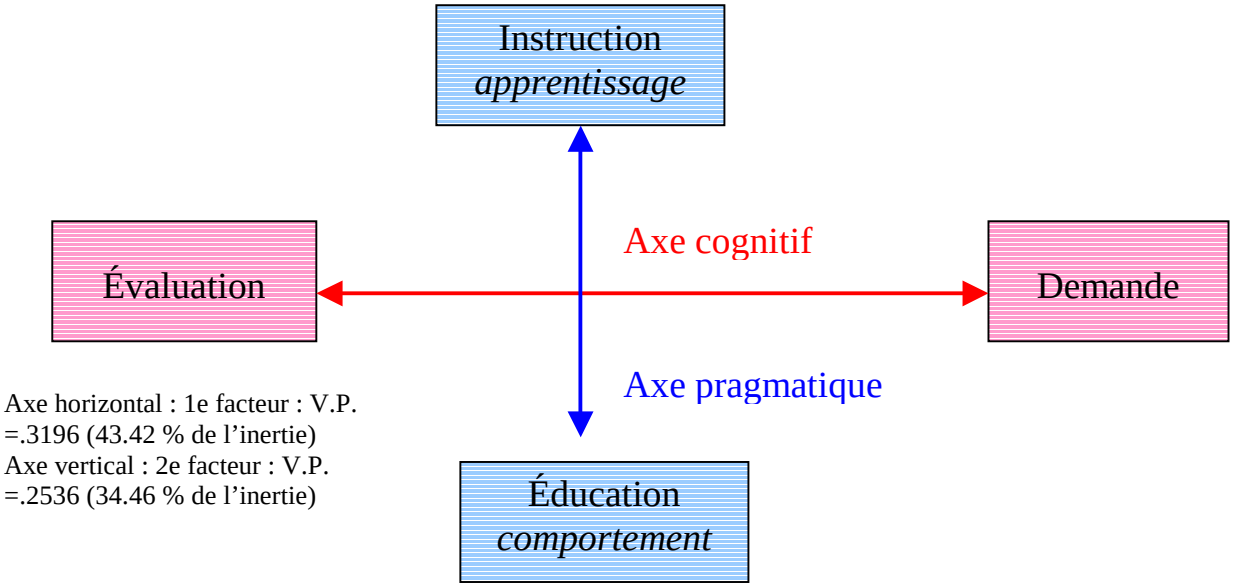
Habituellement, elle donne lieu à une analyse de contenu, parfois à une Analyse de Discours par l'étude des modalisations. Nous allons pousser l'interprétation d'Alceste un peu plus loin grâce aux variables tensives. L'hypothèse de travail sera ici l'interprétation de l'espace vide entre les classes 1 et 2 comme un discours absent ou rarement actualisé, celui de la difficulté d'enseigner.

L'analyse factorielle de correspondances produit le tableau suivant. Comme nous n'obtenons que quatre classes, l'A.F.C. ne crée que qu'une seule projection avec deux axes.

2.3. *Interprétation des axes*

L'A.F.C. propose une représentation spatiale des oppositions entre classes. À partir de l'analyse de discours précédente, nous interprétons les axes d'après le contenu :

Figure 2. Reconstitution des tensions au niveau du contenu

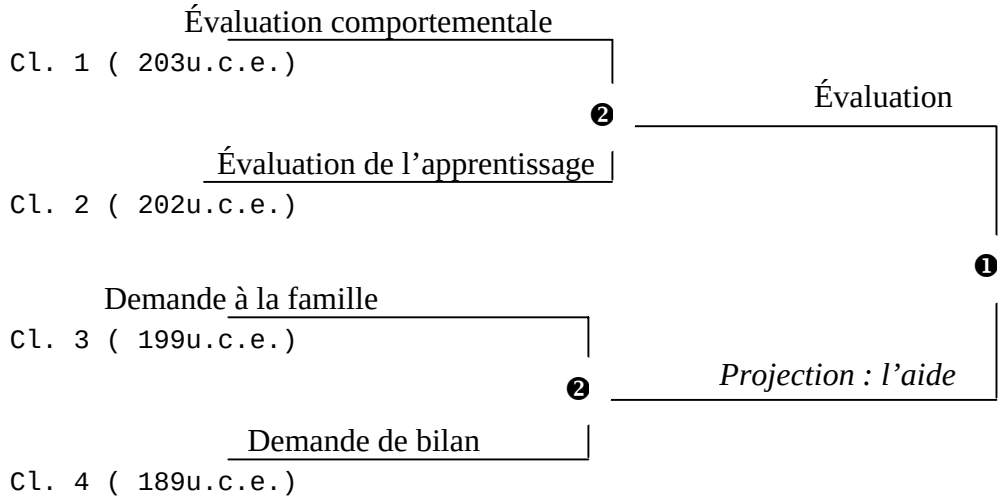


L'axe le plus significatif est l'axe horizontal. Il marque l'opposition entre l'évaluation d'une difficulté actuelle d'une part et d'autre part, la projection dans un futur d'aide ou d'échec. Ensuite, l'axe vertical marque la tension entre un objectif éducatif visant un comportement d'élève et un objectif d'instruction visant des performances cognitives.

Dans une A.F.C., les axes matérialisent des variables internes au corpus. Nous avons ici **deux variables** à nommer, ce qui est difficile. Dans l'état actuel du travail, nous proposons de nommer l'axe horizontal « axe cognitif » parce qu'il distend le corpus entre la demande et l'argumentation à cette demande. L'évaluation est une argumentation pour convaincre le destinataire d'aider l'élève. Nous nommerons l'axe vertical « axe pragmatique » parce qu'il matérialise la tension entre deux états de l'élève « être enfant » ou « être élève ». Mais ces étiquettes seront peut-être modifiées. Ceci nous permet d'attirer l'attention du nouvel utilisateur d'Alceste sur un aspect de l'interprétation. C'est un travail laborieux fait d'aller-retour entre le rapport et le texte, entre les hypothèses qui surgissent et leur vérification. L'interprétation d'un rapport est donc un document qui évolue progressivement. Ce travail nous enseigne l'humilité devant une tâche jamais terminée !

Les axes rejoignent la construction de la classification. Nous pouvons donc les mettre en parallèle avec la Classification Descendante Hiérarchique et ainsi, nommer chaque branche du dendogramme.

Figure 3. Arborescence des classes



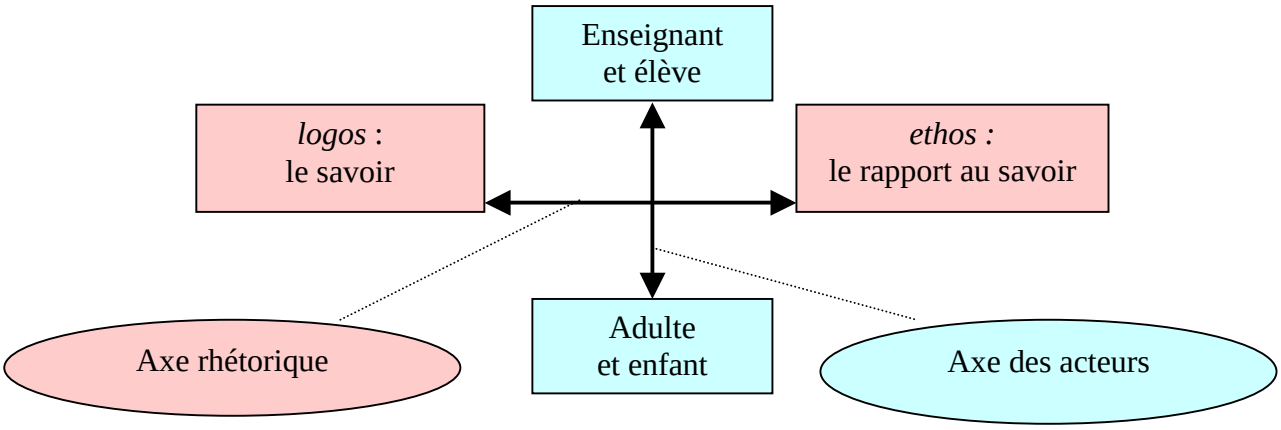
La première partition ❶ correspond alors à l'axe horizontal. La tension la plus significative se fait entre l'évaluation d'une difficulté et la projection d'une aide. Ensuite, logiquement, la deuxième partition ❷ donne naissance à l'axe vertical. Les deux branches mères se divisent donc chacune en deux branches déclinant les deux valeurs mères en sous-classes comportant soit le trait /comportement/ soit le trait /apprentissage/.

3. Au plus profond du texte, les variables tensives

Nous pouvons interpréter les axes de différentes manières. Le chercheur pourra adapter le langage utilisé pour décrire à la problématique qu'il explore et aux théories qu'il convoque. Par exemple, nous avons questionné l'axe horizontal en lien avec **la rhétorique.**

3.1. *Une variable rhétorique : du logos à l'ethos*

Figure 4. Reconstitution des tensions dans la demande d'aide



Le savoir sur l'élève (logos) est utilisé comme argument de la demande (rapport au savoir, la place du sujet interprétatif)

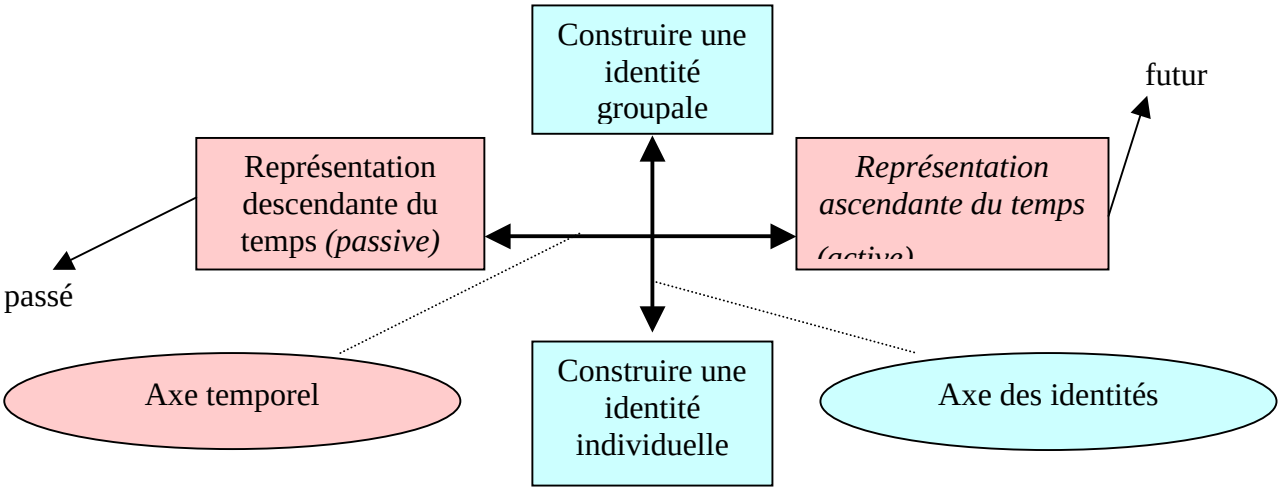
Nous pouvons qualifier cet axe de rhétorique parce qu'il articule un **procès d'argumentation**.

Dans la rhétorique **d'Aristote,** l'*exemplum* est un récit offert pour argumenter, de la même manière qu'une parabole. Ici, le **récit** de la difficulté est mis en scène (gauche de l'axe horizontal) pour servir l'objectif de la demande d'aide (droite de l'axe horizontal). Si Jordan (cas n°28) est décrit ainsi « Jordan a un comportement anormal en classe. Il se montre agité, perturbateur, toujours désireux d'attirer l'attention. », c'est pour donner des informations sur son comportement mais surtout pour argumenter l'aide future.

Ensuite, l'axe horizontal sera interprété **par la praxématique**, c'est-à-dire en rapport avec la construction de l'objet d'étude, des écrits institutionnels à visée pragmatique.

3.2. *Une variable praxématique : représentation du temps*

Figure 5. Reconstitution praxématique de la demande d'aide (cf. Brès, 1994)



Les acteurs envisagés en termes de procès seront décomposés par rapport à leurs différentes identités. Dans la scène de l'école, ils jouent à construire ou déconstruire leurs identités individuelles ou groupales (enfant ou élève – adulte ou enseignant …).

L'axe horizontal reconstitue l'axe temporel du contexte par rapport au point « t » qui est le jour de la demande (repéré puisque la fiche est datée). Ainsi, le récit de la difficulté représente un présent ou un passé. Jacques Brès parle de représentation descendante du temps. Par exemple, le cas 185 décrit le comportement de la fillette entre le 26 avril et le 21 mai, date de la demande.

Extrait du cas 185 : Elle fait des crises de nerf : quatre grosses depuis qu'elle est là : elle se roule par terre, hurle puis reste prostrée sous un banc demi-heure à trois-quarts d'heure. (dernière en date : vendredi 18 mai). Elle ne fait rien, ne veut pas écrire et dès qu'on la bouscule physiquement (exemple : dans la cour de l'école) ou moralement (on lui demande de faire un travail) elle boude puis elle crise.

Nous voyons que le récit du comportement est ancré sur la période de référence soit dans un passé proche (vendredi 18 mai) soit dans un présent (elle ne fait rien). Les performances de la fillette sont toutes au présent. Ce récit décrit les crises, donc des événements qui se sont reproduits plusieurs fois mais qui n'ont pas un caractère permanent.

La représentation du temps est ici une construction du temps passé comme une suite de crises entre le passé et le présent.

Ces deux interprétations n'ont pas la prétention d'être exhaustives. Elles ont deux lectures possibles parmi celles que d'autres interprétants pourraient extraire du texte selon leurs cadres théoriques.

4. Interprétations dynamiques des discours

4.1. *Cendrillon articule le passé et le futur*

La description du comportement de Cendrillon, le cas 29, est suivie d'une demande précise.

1. Cendrillon **a un comportement agité** et **n'arrive pas à fixer son attention**, de plus, **a des difficultés** scolaires.

2. **Dérange la classe** en faisant du bruit, en interpellant ses voisins, en se déplaçant dans la classe trop souvent.

6. X a fait beaucoup d'efforts pour apprendre à lire et voudrait bien faire.

8. Un **soutien scolaire** (**compréhension des consignes** ...)

Le passé est une construction sur l'évaluation du comportement et de l'apprentissage de l'élève. (Cendrillon a un comportement agité et n'arrive pas à fixer son attention, de plus, a des difficultés scolaires. Dérange la classe en faisant du bruit, en interpellant ses voisins, en se déplaçant dans la classe trop souvent.)

Le futur est une demande formulée concrètement. (Un soutien scolaire - compréhension des consignes ...) Nous pouvons parler de représentation active du temps dans le sens où l'enseignante propose un projet élaboré, une action de remédiation.

Le passé est une représentation passive dans le sens où le seul rôle de l'énonciateur est d'agencer le récit. Le matériau qui constitue le récit est contraint par la vie de la classe. La seule part active de l'énonciateur est constituée par l'énonciation, des choix de mise en scène du récit. Par contre, le futur est une représentation active parce qu'il inclut tous les possibles

d'une aide imaginable. La variable tensive « représentation temporelle » trouve ici une actualisation entre un passé de difficulté scolaire et un futur d'aide.

4.2. *Jules et sa maîtresse ont besoin d'aide*

Nous avons sélectionné le cas de Jules présenté par Madame Dupont, cas dont la description articule nettement la classe 2 (2 u.c.e.) et la classe 4 (4 u.c.e.).

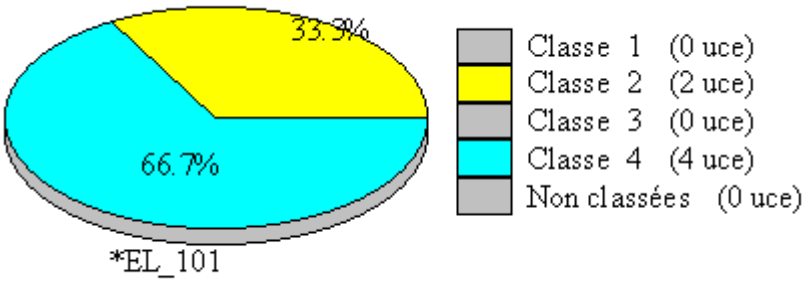


Figure 6. Distribution des u.c.e. dans les classes pour le cas 101

Comment le contenu est-il réparti entre ces deux classes ?

cas 101

1. Après deux CP, Jules a encore de grosses difficultés notamment en lecture. Il a du mal à déchiffrer, inverse les lettres, confond les sons. Il entend mal les sons (problème d'oreilles ?). Difficultés pour s'exprimer clairement (problèmes de syntaxe, de vocabulaire et de prononciation). Problèmes de compréhension. Problèmes familiaux.

2. Problème d'orientation pour l'année prochaine. Je ne suis pas certaine que son année de CE1 lui sera bénéfique.

3. Dans la classe, travail individualisé : méthode de lecture CP (il ne participe pas à certaines matières notamment en langue française).

4. VIDE

5. Il allait chez l'orthophoniste l'année dernière mais les séances ont été interrompues pour une raison un peu floue à mon avis (d'après la vision de la mère).

6. C'est un enfant qui a de la volonté notamment pour apprendre à lire.

7. OUI. Le père aide Jules à faire ses devoirs (lecture) en commun accord avec moi. Le problème est que je n'ai rencontré que la maman avec qui le dialogue n'est pas toujours facile.

8. Je souhaiterais un bilan scolaire (et autre ?) pour m'aider à adapter ma pédagogie et pour l'orientation l'année prochaine.

Le corpus est décomposé ainsi :

- Des u.c.e. constitutifs de la classe 2 :

```
448  20  1. apres deux CP, X a encore de #grosses #difficultes #notamment
en #lecture. il a du mal a #dechiffrer, inverse les #lettres, confond les
#sons.
449  20  il #entend mal les #sons, #probleme d' oreilles. #difficultes pour
s' #exprimer clairement, #problemes de syntaxe, de #vocabulaire et de #pro-
nonciation.
```

L'évaluation cognitive de Jules est dite avec grosses *difficultés* notamment en lecture.

Il a du mal à *déchiffrer*, inverse les lettres, confond les sons. Il a des *problèmes* avec la langue française.

- Des u.c.e. constitutifs de la classe 4 :

```
447  39  8. je #souhaiterais un #bilan scolaire et ou #psychologique pour
#savoir quel #type d' #aide il faut lui #apporter£
454  33  8. je #souhaiterais un #bilan scolaire, et autre? pour m' #aider a
#adapter ma #pedagogie et pour l' #orientation l' #annee #prochaine£
```

Les aides demandées sont un *bilan scolaire* ou *psychologique* pour Jules et pour Madame Dupont, des conseils *pédagogiques* ainsi que des conseils pour *l'orientation.*

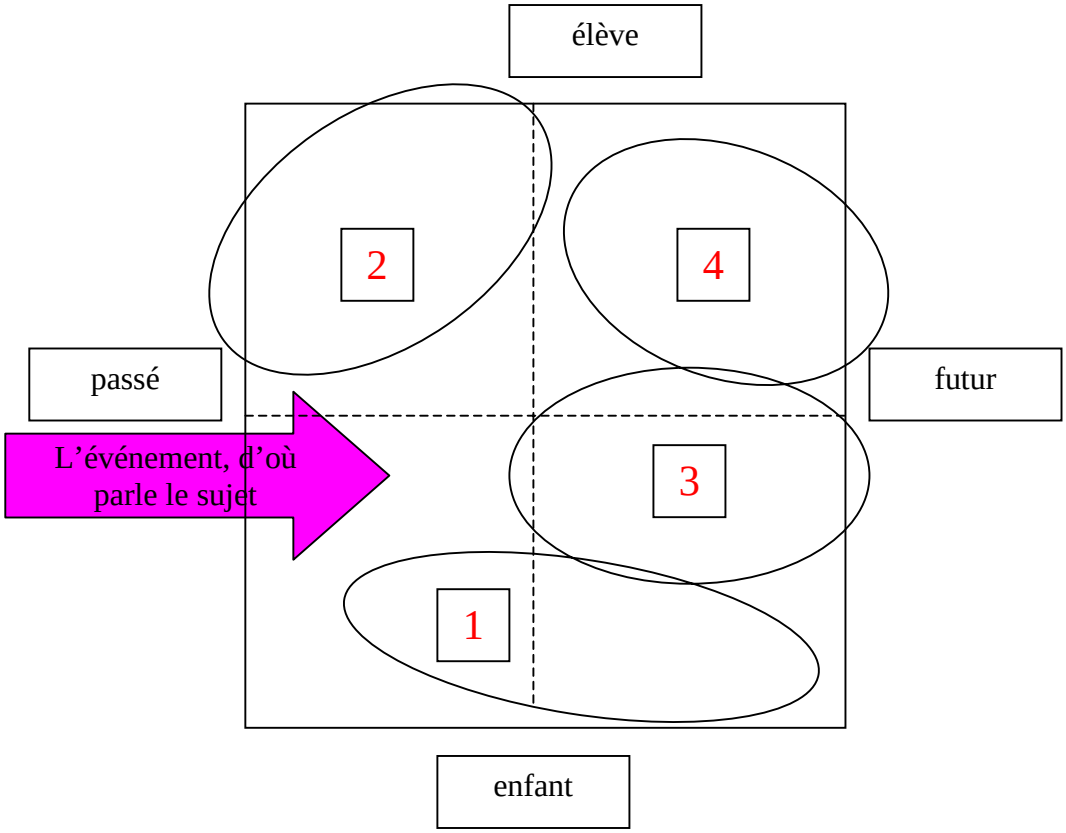
Ce cas illustre pleinement l'axe cognitif, qu'il soit défini :

- dans son contenu comme **cognitif**, de l'évaluation à la demande,
- dans son discours comme **objet rhétorique**, du logos à l'ethos,
- dans son discours comme **représentation temporelle**, du passé de difficulté au futur d'aide.

4.3. *Des mouvements entre les discours*

Dans les paragraphes précédents, nous avons discuté des discours articulant les classes selon différentes interprétations de l'axe horizontal. Pour interpréter ces variables représentées par les axes, nous allons tenter de revisiter la **tension** (cf. Fontanille, 1998). Le plan orthonormé formalisant les résultats de l'A.F.C. sera considéré comme un espace au même titre que l'espace physique d'une ville ou d'un autre espace géographique. Les centres attracteurs ne sont pas des lieux de vie mais des centres de classes, donc des centres de discours. L'attraction se fait ici par des types de discours, la typologie étant considérée comme la présentation d'un modèle social qui aide à la parole. Une analyse sémiotique (qui ne sera pas développée ici) nous a orienté vers des analyses comparatives.

Figure 7. Une interprétation du discours absent



L'espace entre la classe 1 et la classe 2 ne peut manquer d'attirer notre attention. L'A.F.C., qui matérialise les tensions entre mondes lexicaux, nous montre un vide (cf. fig.1). Nous allons poser l'hypothèse d'une irrégularité signifiante. L'événement aurait produit l'absence, aurait créé un espace de signification non actualisé. Le « discours absent » bousculerait donc l'ordonnancement logique des discours. Par une première observation de l'A.F.C., nous pouvons dire que l'évaluation comportementale (classe 1) s'éloigne de l'élève pour un futur du côté de la famille (classe 3). La demande à la famille (classe 3) s'éloigne légèrement de l'enfant pour s'orienter vers l'élève futur.

Le « discours absent » est seulement actualisé par deux ou trois collègues (dont Madame N) et marqué par les mots *travail* et *comprendre* (cf. fig.1). Il apparaît dans l'énonciation, entre autres, par des énoncés négatifs[4] et le pronom indéfini *on*[5].

5. Le discours absent, un hiatus entre les variables

Que signifie ce vide ? (entre la classe 1 et la classe 2). Les seuls mots occupant cet espace sont *travail* et *comprendre.* La représentation montre un hiatus du 2ème facteur. Quel est le statut de ces deux mots qui n'appartiennent ni à la classe 1, ni à la classe 2 ou aux deux ?

5.1. *Parler du travail*

Cet entre-deux signifie-t-il une appartenance aux deux classes ?

```
Classe 1 :  534    36. 109.   33.03    3.66      Y travail<

Classe 2 :  534    40. 109.   36.70    8.39      Y travail<

Vocabulaire spécifique de la classe 2 :

B4 travail< : travail(37), travaille(3), travailler(5);
```

Nous voyons que le **nom travail** (37 occurrences) est massivement préféré au **verbe travailler** (5 occurrences). Les situations parlent donc d'un produit, le travail plutôt que d'un procès le verbe travailler.

Ensuite, dans la répartition des classes, il est présent à la fois dans la classe 1 (Évaluation comportementale) et la classe 2 (Évaluation cognitive), avec toutefois une différence de Khi2. Il est associé plus spécifiquement à la classe 2. La représentation graphique de l'A.F.C. le situe dans l'espace entre 1 et 2, prend donc en compte l'appartenance aux deux discours.

Si le nom « travail » est relié à la fois aux classes 1 et 2, c'est peut-être pour cette raison qu'il est situé dans l'entre-deux.

5.2. *Parler de la* compréhension *ou de* comprendre

Le verbe « comprendre » se situe aussi dans l'entre-deux, entre la classe 1 et la classe 2 alors que le nom « compréhension » est très lié à la classe 2 (Khi2 60,12).

```
Classe 2    334    24.  56.   42.86    9.59      V comprendre.

Classe 2  B4 comprendre. : comprend(12), comprendre(12), comprenne(1);
```

La vérification par les Khi2 montre qu'il est tout de même très lié à la classe 2.

[4] Cf. communication au colloque Sémiologie 2003, laboratoire Dynalang, Paris V, Sorbonne, novembre 2003. Actes à paraître.

[5] Cf. communication au congrès de l'Association Française de Sociologie à Villetaneuse en février 2004. Actes à paraître.

```
Classe 2 : 465    30.  38.   78.95   60.12      Y comprehensi<

Classe 2 : B9 comprehensi< : compréhensible(1), compréhension(29);
```

Cette distinction nous éclaire sur l'un des aspects du discours absent. Le nom « compréhension » est très lié à la classe 2 alors que **le verbe « comprendre » appartient au discours « absent ».** La *compréhension* est un **objet** que l'on peut jauger, évaluer, caractériser en fonction de son degré de jonction avec le sujet apprenant (bonne compréhension ou compréhension limitée …). Par contre, *comprendre* est un **procès** qui relie deux états de jonction le sujet apprenant incompétent et le sujet apprenant compétent. Au cours du procès « comprendre », se joue toute la scène de l'école incluant l'enseignant, l'élève, les autres élèves …

5.3. *Le discours absent, la difficulté d'enseigner*

Qui est cet enfant qui comprend ou ne comprend pas, qui réalise un travail ou ne le réalise pas et dont on ne parle pas, ou si peu ? Le comportement de l'enfant se dit parfois (dans la classe 1) mais entre l'évaluation du comportement et l'évaluation cognitive, il ne semble pas y avoir de point commun, sinon par les deux seuls mots actualisés : *comprendre* et *travail*.

Notre interprétation pragmatique de ce discours absent met l'accent sur **la pédagogie**. Le comportement de l'enfant doit d'abord être scolaire (le métier d'élève) pour que le cognitif s'enclenche. En d'autres mots, en classe, un enfant doit d'abord devenir élève pour réunir les conditions nécessaires à l'apprentissage. Entre les deux classes d'Alceste, nous pouvons parler de **conditions nécessaires à l'apprentissage**, donc de relation pédagogique. Pourquoi ce discours n'est-il pas actualisé ?

Dans le cas de Jules, le récit met en scène un élève en difficulté. La demande de Madame Dupont concerne à la fois l'élève (bilan scolaire ou psychologique) et l'enseignante (des conseils pédagogiques ainsi que des conseils pour l'orientation). Si L'énonciateur demande de l'aide pour l'enseignante, nous pouvons présupposer qu'elle ressent des difficultés dans ses différentes missions. Le discours absent défini par « comprendre » présuppose la relation pédagogique, la difficulté de l'enseignante à assumer sa mission soit *faire comprendre à Jules*. Entre l'enseignante et l'énonciatrice (deux rôles assumés par le même acteur, Madame Dupont), la difficulté d'enseigner est mise en scène, occultée pour ne laisser apparaître que l'élève dans une relation pédagogique difficile.

Le discours absent est intriqué entre deux variables tensives, la variable « rhétorique » et la variable « acteurs ». Entre le récit et la demande, se cache la présentation d'un acteur (l'élève) et une non-présentation d'un autre acteur (l'enseignant). Cette mise en scène est l'une des composantes du *logos* de la demande d'aide, la difficulté d'enseigner indissociable de la difficulté d'apprendre dans l'école mais dissociés par la rhétorique.

Conclusion

Les variables tensives, qu'elles soit nommées *rhétorique* ou *temporelle*, donnent un contenu aux discours actualisés. Le cas de Cendrillon articule un passé de difficulté comportementale (comportement agité, n'arrive pas à fixer son attention) à un futur d'aide projetée (un soutien scolaire, compréhension des consignes). Jules est présenté dans un passé de difficultés cognitives (grosses *difficultés* en lecture, Il a des *problèmes* avec la langue française) justifiant une aide future dans deux directions : un *bilan scolaire* ou *psychologique* pour Jules et pour Madame Dupont, des conseils *pédagogiques* ainsi que des conseils pour *l'orientation*. Ce qui tend à faire apparaître une absence, la place de l'enseignante dans la difficulté présentée. Or, les nuages des *topoï* montrent un vide, le discours absent. Ce discours serait l'événement d'où

parle le sujet, la difficulté d'enseigner, l'une des parties du procès de l'apprentissage. Si Jules n'apprend pas, l'enseignante est en échec dans cette mission. Ce contenu apparaît en creux entre les variables tensives, par un choix de présentation des acteurs (axe vertical) actualisés ou pas entre *logos* et *ethos*.

L'utilisation des variables tensives semble donc opérant pour une interprétation dynamique des contenus et des discours

Références

Bres J. (1994). *La narrativité.* Duculot, coll. Champs linguistiques.

Fontanille J. (1998). *Sémiotique du discours*. PULIM.

Jenny J. (1997). Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. État des lieux et essai de classification. *Bulletin de Méthodologie sociologique*, vol. (54) : 41-64.

Jenny J. Pour engager un débat avec Max Reinert, à propos de la réponse "Alceste" à l'"objet" d'une analyse de discours de type statistique. *Langage et Société*, vol. (90) : 73-85.

Reinert M. (1990). Alceste une méthode d'analyse de données textuelles et une application : *Aurélia* de Gérard de Nerval. *Bulletin de méthodologie sociologique*, vol. (26) : 24-54.

Reinert M. (1997). Les « mondes lexicaux » des six numéros de la revue *Le surréalisme au service de la révolution*. In *Mélusine - cahiers du Centre de Recherche sur le Surréalisme*, vol. (16) : 270-302.

Reinert M. (1998). Mondes lexicaux et *topoï* dans l'approche *Alceste.* In Mellet S. et Vuillaume M. (textes rassemblés par), *Mots chiffrés et déchiffrés – Mélanges offerts à Etienne Brunet*. Honoré Champion.

Reinert M. (1999). Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "Alceste". *Langage et société*, vol. (90) : 57-70.

Reinert M. (2001a). Processus catégorique et co-construction des sujets et des mondes à travers l'analyse statistique de différents corpus. In Arrivé M. et Normand C., *Linguistique et Psychanalyse* [colloque de Cerisy, sept. 1998]. Éditions In Press : 379-392.

Reinert M. (2001b). Alceste une méthode statistique et sémiotique d'analyse de discours – Application aux "rêveries du promeneur solitaire". *Revue française de psychiatrie et de psychologie*, vol. (49 : V).

Reinert M. (2003). Le rôle de la répétition dans la représentation du sens et son approche statistique par la méthode Alceste (à paraître).

Wald P. (1999). Classes d'énoncés, dimensions modales et catégorisations sociales dans ALCESTE. *Revue de sociologie et d'anthropologie Utinam*, vol. (1/2). L'Harmattan.

Évaluation d'outils de Text Mining : démarche et résultats

Yasmina Quatrain[1], Sylvaine Nugier[1], Anne Peradotto[1], Damien Garrouste[2]

[1]EDF R&D – 92141 Clamart Cedex – France
[2]Lincoln – 92774 Boulogne-Billancourt Cedex – France)
yasmina.quatrain@edf.fr, anne.peradotto@edf.fr

Abstract

Électricité de France, the French electricity provider is moving into new markets, and needs to know very well its clients and tracks their satisfaction. Thus, different kinds of text mining tools were tested in order to analyze a huge quantity of heterogeneous textual documents, including mails, open-ended customer survey questions, discussion forums, and comments contained in large databases concerning customer contacts. In this context, it seemed unavoidable to build a test grid, in order to facilitate the comparison between different tools. This grid was created according to a test on data mining software, achieved by a consulting firm. However, we tried to give equal importance to the statistics and the linguistics aspects. It is divided in ten points, ranging from the company which developed the tested software to the methods implemented and data extraction. This paper describes the test grid carrying out, the software selection, the way the evaluation of four tools (Alceste, SAS Text Miner, TEMIS Insight Discoverer and SPAD/CRM) was achieved and the results. In conclusion, this work underlines that text mining tools are of two types: statistics software including linguistics modules and linguistics software using statistical methods to analyze text. Furthermore, the evaluation stressed on the obligation of using detailed test grids and creating evaluation protocols, in order to define the functionalities which are essential to be tested in a software, according to the user's aims and profile and the kind of corpora he or she would like to analyze.

Résumé

Dans le contexte de l'ouverture du marché de l'électricité, EDF désire analyser les gros volumes de données textuelles qui lui permettront de mieux connaître ses clients. Dans cette optique, plusieurs outils de text mining destinés à l'analyse de cette information hétérogène de taille importante ont fait l'objet d'une évaluation à l'aide de trois corpus de nature différente. La constitution d'une grille de test facilitant la comparaison des logiciels est apparue indispensable. Inspirée d'une expertise sur les outils de data mining, elle a été réalisée en évitant de privilégier la communauté statistique au détriment de la linguistique. Cette grille se compose de dix thèmes variant de la société éditrice aux champs d'application en passant par l'accès aux données et l'analyse du tableau lexical. Outre le déroulement de l'évaluation et ses résultats sur quatre outils du marché (Alceste, SAS Text Miner, TEMIS Insight Discoverer et SPAD/CRM), cet article retrace la démarche de constitution de la grille de test, le choix des outils évalués et les critères retenus. Il conclut sur l'existence de deux types d'outils en text mining : ceux statistiques enrichis d'un module de traitement du texte et ceux originellement linguistiques. De plus, cette expérience conforte l'utilisation d'un protocole détaillé permettant de déterminer et évaluer les fonctionnalités incontournables en fonction des objectifs et du profil de l'utilisateur du logiciel et de la nature du corpus à analyser.

Mots-clés : évaluation, protocole, text mining, logiciels.

1. Introduction

Le volume croissant de données textuelles provenant de l'Internet et des contacts clients (par mails, retranscription de message téléphonique, lettres de réclamation, enquête…) apportent une quantité d'informations non exploitable manuellement et actuellement peu exploitée à

Électricité De France (EDF). Pourtant ces données sont indispensables à une bonne connaissance des clients et à l'amélioration de la gestion de la relation avec celui-ci, notamment dans le contexte actuel d'ouverture à la concurrence.

Le processus d'extraction de cette information à partir d'un large volume de données textuelles non structurées est appelé text mining. Il débute par une étape primordiale de préparation des données allant de la structuration à l'enrichissement en passant par le filtrage. Les méthodes employées ensuite pour extraire l'information pertinente sont principalement des tâches de classification et de classement.

C'est dans ce contexte que nous avons été confrontés au choix d'un logiciel traitant de données textuelles (text mining), pour différents types d'études. Par outils de text mining, nous entendons ici des outils permettant l'analyse de données non-structurées (textuelles dans notre cas) associées à des données structurées telles que les données relatives à la consommation d'un client ou à son type de logement. Afin d'aider à l'évaluation et à la comparaison de ces outils, la constitution d'une grille de test est apparue nécessaire.

Cette grille avait ainsi deux principaux objectifs :
– Permettre l'évaluation comparée de logiciels analysant des données textuelles ;
– Être une aide à la décision pour l'achat d'un tel logiciel adapté au besoin des utilisateurs.

La démarche s'est voulue pragmatique en recherchant tout d'abord les expériences similaires dans le domaine du data mining et du traitement du langage naturel.

Une évaluation de logiciels de data mining (CXP, 2001) a permis d'avoir un bon aperçu des volets devant figurer dans notre grille. Il ne traitait pas la partie relative aux fonctionnalités de traitement du texte, inexistante dans ce domaine. Nous avons trouvé assez peu de références sur la réalisation de protocoles d'évaluation ou l'évaluation d'outils de cette nature (Brugidou *et al.*, 2000). Les articles traitant du sujet concernent des logiciels très spécifiques tels que l'évaluation pour le résumé automatique (Barthel *et al.*, 2002) ou celle d'analyseurs syntaxiques (Aït Mokhtar *et al.*, 2003). Le projet européen TECHNOLANGUE, co-financé par le Ministère français délégué à la recherche et aux nouvelles technologies, le Ministère français délégué à l'Industrie et le Ministère de la Culture et de la Communication, comporte un volet évaluation, décliné en domaines de traitement du langage naturel, dans lesquels ne figure pas le text mining (dont EVALDA-ARCADE II pour l'évaluation de l'alignement de corpus multilingues ou EVALDA-CESTA pour les systèmes de traduction automatique).

La grille devant permettre l'évaluation de nos outils a été construite à l'aide des références précédemment citées, et a évolué tout au long des tests sur les trois types de corpus retenus (questions ouvertes d'enquête, champs commentaires d'une base de données sur les contacts Clients, forums de discussion).

Nous avons finalement retenu un ensemble de critères organisés en trois axes :

– le commercial (prix, prestations, documentation…) ;
– le technique (architecture, limites volumétriques…) ;
– le fonctionnel (traitements possibles, convivialité, exploitabilité des résultats...).

Cet article contient la démarche, la grille de test appelée PITT (grille Pour l'évaluation de logiciels statistiques Incluant le Traitement du Texte) et les résultats sur les quatre outils évalués. Un objectif à plus long terme est la création d'un véritable protocole de test d'outils de text mining.

2. Élaboration de la grille de test

2.1. *La démarche*

Au cours de la construction de la grille de test, nous avons cherché à éviter de privilégier la communauté statistique au détriment de celle du traitement automatique des langues, l'une plus sensible aux moyens mis à disposition pour l'analyse du texte, l'autre souhaitant trouver des méthodes élaborées d'analyse de données ou de modélisation. De plus, nous ne souhaitions pas lister plus ou moins exhaustivement les fonctionnalités disponibles dans un ensemble d'outils sans se pencher plus particulièrement sur les méthodes applicables au texte.

Pour le premier point il convient d'utiliser un langage compréhensible par les deux communautés. Pour le second, et à l'inverse des outils de data mining qui offrent des gammes de plus en plus larges de méthodes, il semble important de recentrer les études à partir des propriétés essentielles (capacité à classifier, capacité à classer) que l'on attend d'un outil de text mining et dont les résultats sont interprétables et pertinents.

2.2. *Le choix des logiciels évalués*

Les logiciels testés ne forment pas une liste exhaustive de l'offre du marché, cependant le choix s'est porté volontairement sur des logiciels différenciés offrant ainsi une large couverture technique (dans les méthodologies proposées) et une large couverture opérationnelle (dans les applications possibles).

Le choix des logiciels s'est fait en deux temps : SAS étant l'outil statistique de référence à EDF, le test du nouveau module SAS/Text Miner est apparu évident. De même pour Image/Alceste, outil utilisé depuis plusieurs années à la R&D d'EDF pour l'analyse de textes. Les deux autres outils ont alors été choisis pour compléter la gamme avec comme objectif de balayer au maximum l'ensemble des fonctionnalités offertes par ces outils dits de text mining. Nous avons ainsi sélectionné la suite TEMIS Insight Discoverer présentée comme une véritable solution de text mining, et un logiciel de statistique permettant l'analyse du texte, SPAD/CRM de la société DECISIA.

Ces quatre logiciels donnent donc un très bon aperçu des outils existants dans le paysage du text mining pour les raisons suivantes :

– Les orientations « commerciales » sont différentes.

Alceste est un logiciel dédié à l'analyse de données textuelles utilisé pour traitement du discours sans fonctionnalités inférentielles. SAS Text Miner est une solution de text mining intégrée dans une suite logicielle de data mining. SPAD/CRM se positionne également sur le marché du data mining. TEMIS Insight Discoverer est un outil exclusivement text mining.

– Les méthodologies et fonctionnalités adoptées sont larges.

On observe des différences très importantes dans la partie du traitement purement textuel (présence ou non d'un outil linguistique plus ou moins performant), dans la construction et la réduction des tableaux lexicaux (analyses factorielles ou autres) et enfin dans les méthodes d'analyse proposées.

– Les degrés de maturité sont différents.

Alceste et SPAD/CRM (anciennement appelé SPAD-T) sont des logiciels éprouvés dans l'analyse de données textuelles, la solution SAS Text Miner est par contre très récente (la version que nous testons de SAS est la première version commercialisée), de même que la solution TEMIS Insight Discoverer.

2.3. *Le déroulement du test*

Il est important que les outils de text mining soient tous testés selon le même principe afin de garantir la comparabilité des résultats et sa relative pérennité dans le temps. Les tests ont été effectués à partir d'une machine unique et par la même personne, dans le cadre d'un stage de licence professionnelle en data mining.

Pour chacun des corpus, les fonctionnalités de chaque logiciel ont été éprouvées en renseignant au fur et à mesure les différents points de la grille de test.

Les corpus choisis pour les tests représentent un échantillon non représentatif de l'ensemble des données textuelles pouvant être analysées par les méthodes de text mining. Cependant ils ont été sélectionnés pour leurs caractères différenciés et parce qu'ils correspondent aux différents types de corpus que nous avons été amenés à analyser. Ces derniers sont les suivants :
– Corpus « QO » : réponses à une question ouverte dans une enquête de satisfaction EDF ;
– Corpus « commentaires » : champs commentaire extrait d'une base de données EDF ; ce champ est renseigné, en cas de nécessité, par l'agent à la suite de contacts téléphoniques avec les clients ;
– Corpus « forums » : forums de discussions sur internet.

La nature des données (langage naturel ou retranscription) est différente suivant nos corpus : Les corpus « QO » et « forums » sont du langage naturel contrairement à celui des « commentaires » où les motifs d'appels des clients ont été retranscrits en abréviations par un opérateur.

La volumétrie est également très variable d'un corpus à l'autre : Le corpus « commentaires » est constitué de 100 000 motifs d'appel, « forums » contient 2 000 interventions et « QO » 400 réponses à une question ouverte sur la satisfaction.

2.4. *Les critères retenus*

Les critères retenus se déclinent en 10 thèmes, chaque thème contenant des sous-thèmes non détaillés dans ce chapitre.

1. La Société

Cette rubrique a pour but d'indiquer l'origine de l'outil. Cela permet de déterminer si ce dernier suit une école en particulier (statistique à la française ou autre). Par le biais du pays d'origine, cette rubrique renseigne aussi sur la langue pour laquelle le produit a été conçu. Les autres entrées permettent de déterminer la santé financière de la société à l'origine du logiciel, la pérennité de ce dernier et son potentiel d'évolution au sens financier.

Items évalués :
Origine – Nom de la Société – Effectif France/Monde – CA France/Monde - Gamme de produits – Existence Club utilisateurs – Site Web.

2. Le Produit – Aspects Financiers

Cette partie concerne le coût de la licence et permet ainsi de définir si un déploiement de large envergure nécessite des moyens financiers importants. Les aspects formation et conseil (prestation) permettent d'évaluer la durée de mise en place d'une première étude.

Items évalués :
Achat/location – Coût Licence – Coût d'assistance / maintenance évolutive – Formation – Existence de partenaires en conseil et développement sur le produit.

3. Le Produit – Aspects Techniques

Trois sous-parties composent ce thème : l'architecture, la prise en main et les généralités.

Les items relatifs à l'architecture doivent permettre de répondre à des questions du type : Quelle architecture faut-il mettre en place ? Le produit est-il adapté à l'environnement de travail de l'utilisateur ? S'insert-t-il dans l'environnement technique en vigueur dans l'entreprise ? L'achat du logiciel nécessite-t-il l'achat de machines supplémentaires plus récentes ? …

Les items relatifs à la prise en main nous renseignent sur la facilité avec laquelle l'utilisateur s'approprie l'outil. Elles font également l'inventaire des supports à sa disposition.

Des remarques générales sur le logiciel, apportant des détails sur l'appréhension par l'utilisateur de sa philosophie, sa fiabilité et la manière dont les erreurs sont gérées sont réunies sous la dernière rubrique.

Items évalués :
Mode Client/Serveur – Systèmes d'exploitation supportés – Ressources mémoire nécessaires – Espace disque nécessaire – Logiciels nécessaires – Historique des versions
Niveau de l'utilisateur – Existence d'aide en ligne, manuel d'utilisation – Documentation Méthodologique
Niveaux de fiabilité (fréquence des bugs) – Présence d'un système de gestion des erreurs (traces) – Présence d'une organisation en mode projet (organisation des travaux) – Degré d'ouverture et de personnalisation de l'outil – Existence de sauvegarde des paramétrages et résultats – Appréciation générale sur le temps de calcul – Appréciation sur la transparence des méthodes statistiques utilisées – Appréciation générale sur l'ergonomie, la facilité d'apprentissage, la convivialité.

4. L'accès aux données

Cette partie permet de déterminer d'une part la facilité d'intégration des données textuelles et extra-textuelles au logiciel (accès, formatage, langage de programmation spécifique, etc.) et d'autre part ses limites en terme de volumétrie des données pouvant être analysées.

Items évalués :
Possibilité d'accéder à distance aux documents – Duplication obligatoire des documents – Passage par outil d'extraction des données
Formats de documents supportés en entrée (Liste) – Accès direct aux SGBD – Appréciation sur le degré de pré-formatage utilisateur
Nombre d'observations maximums autorisées – Nombre de colonnes maximums autorisées – Nombre de caractères maximums des champs textes – Taille maximum des noms de variables.

5. Les pré-traitements et l'identification des unités textuelles

Les fonctionnalités propres au pré-traitement du texte avant le lancement des analyses sont réunies dans cette section. Cette dernière indique la marge de manœuvre et les facilités offertes à l'utilisateur par le biais de l'interface mise à sa disposition, afin de réaliser les tâches suivantes : nettoyer son texte et le mettre aux normes du produit, l'étiqueter, l'enrichir et enfin visualiser, éditer, importer ou exporter le corpus traité ou le tableau lexical (voir définition au critère 6) construit.

Items évalués :
Pré normalisation des documents (orthographe, harmonisation, majuscules...) – Langues sup-

portées (liste) – Reconnaissance automatique – Mélange de langues – Reconnaissance du rôle grammatical dans la phrase – Reconnaissance des groupes nominaux – Lemmatisation, stemmatisation, autres – Reconnaissances d'entités (noms propres, nombres, adresses,...)
Reconnaissances co-occurrences – Regroupement de termes possible (utilisateur) – Création liste de synonymes – Existence liste de synonymes – Mise à jour liste de synonymes – Création liste de mots vides – Existence liste de mots vides – Utilisation possible d'une liste de départ – Outil de visualisation du corpus sélectionné – Outil d'édition du corpus sélectionné – Statistiques lexicométriques sur le corpus (nombre de termes, d'hapax) – Exportation du tableau lexical entier pour analyses utilisateurs – Importation possible d'un tableau lexical entier (interne ou externe).

6. La transformation et la réduction des tableaux lexicaux

On appelle « tableau lexical » le croisement des unités lexicales (n-grammes, mots lemmatisés, …) et des documents.

Cette rubrique concerne les transformations et réductions possibles du tableau lexical dans l'outil. La transformation de ce tableau aborde notamment les différents types de pondérations des fréquences sur les unités lexicales. La réduction peut être réalisée par filtrages sur ces dernières (sur les fréquences, pondérées ou non, sur un nombre d'unités…), ou par des méthodes plus complexes de type analyses factorielles.

La possibilité de paramétrage du logiciel au niveau de ces transformations est également évoquée, ainsi que la visualisation et la modification manuelle du tableau réduit avant de nouvelles analyses.

Items évalués :
Méthode(s) de transformation des fréquences (richesse, pertinence) – Méthode(s) de réduction des tableaux lexicaux (richesse, pertinence) – Paramétrage possible – Visualisation et d'édition du tableau lexical réduit.

7. L'analyse du tableau lexical

Cette rubrique répertorie les méthodes d'analyse du tableau lexical réduit en les divisant en deux catégories : la classification et le classement, qui sont les deux grands objectifs du text mining. La caractérisation des classes à l'aide de variables non textuelles est également abordée sous cette rubrique.

Items évalués :
Croisement avec des variables non textuelles
Classification : Méthode(s) de classification implémentée (Liste) – Possibilité d'inclure des variables externes actives – Détection automatique du nombre de classes
Autres méthodes
Classement : Liste des méthodes – Méthodes de validation – Scoring.

8. La gestion et la présentation des résultats

Cette section permet d'indiquer la nature de la restitution des résultats par l'outil. Une génération automatique de rapports, l'intégration aisée de graphiques dans des documents préexistants, ou encore l'aide à l'interprétation sont autant d'éléments jouant en la faveur de l'outil et permettant à l'utilisateur de réduire ses délais de traitement.

Items évalués :
Retour au texte initial – Représentations graphiques des résultats statistiques

Éditions d'aides à l'interprétation – Éditions de rapports – Appréciation générale sur la gestion et la présentation.

9. Les champs d'applications

Les champs d'application ont pour objectif général d'aider un futur utilisateur à choisir le logiciel le plus adapté à ses besoins et à ses connaissances. Ils représentent une synthèse des rubriques précédentes, à laquelle s'ajoute l'expérience acquise lors du test des logiciels sur les différents corpus.

Item évalué :
Type d'utilisation conseillée.

10. Les perspectives

Les perspectives servent à estimer la viabilité du logiciel et sa marge d'évolution.

Des outils peuvent être issus de travaux de recherche, du secteur industriel, être plus récents que d'autres, tous ces éléments doivent être renseignés afin de permettre une meilleure appréhension de l'avenir du logiciel.

Items évalués :
Pérennité de l'éditeur – Maturité de l'outil (marge de progression) – Défauts majeurs – Mises à jours prévues.

3. Les résultats et l'interprétation

3.1. *Le test des quatre outils*

Le tableau suivant synthétise les résultats du test des outils Alceste, SPAD/CRM, Text Miner et TEMIS Insight Discoverer[1] :

	ALCESTE	SPAD/CRM	SAS/Text Miner	TEMIS/Insight Discoverer
La Société				
Origine	France	France	USA	France
Pérénité	*	**	***	**
Le produit				
Coût du produit	Très accessible	Moyen	très cher	très cher
Architecture	PC	PC	CL/Serveur	CL/Serveur
Prise en main	Très rapide	Très rapide	Difficile (SEM)	Moyenne
Accès aux données				
formats	*	* *	* * *	***
limites (volumétrie)	*	* *	* * *	***
Pré-traitement				
outils linguistiques	Très complet	Interface pour effectuer manuellement des regroupements et filtrages	Présence de nombreuses fonctionnalités, mais résultats erronés	Très complet
Intervention manuelle	Possible mais inutile	Obligatoire, mais assez conviviale	Obligatoire en l'état actuel du produit mais pas conviviale	Possible mais inutile
Classification				
Diversité des méthodes de réduction du tableau lexical	*	* *	* * *	*

[1] Le détail de l'évaluation est disponible sur demande.

	ALCESTE	SPAD/CRM	SAS/Text Miner	TEMIS/Insight Discoverer
Diversité des méthodes d'analyse des données	* *	* * *	*	*
Classement Méthodes de modélisation	0	* *	* * *	*
Résultats Lecture des résultats Rapport	 * * * * * *	 * * * *	 * *	 *** ***
Champs d'application préconisés	Outil linguistique efficace, Détection rapide des thèmes d'un corpus et bonne caractérisation des classes thématiques	Outil linguistique inexistant, Outil très efficace pour l'analyse des données	Outil linguistique très défaillant, Outil efficace pour effectuer du Data Mining	Outil linguistique et statistique efficace mais absence de caractérisation des classes thématiques

La notation de certaines rubriques varie de médiocre (*) à très satisfaisant (***).

Les évaluations effectuées dégagent un certain nombre d'oppositions et de rapprochements entre les différents outils tant au niveau de l'utilisation, de la méthodologie, des fonctionnalités linguistiques et des fonctionnalités statistiques. Une synthèse a été effectuée afin de pouvoir déterminer les positionnements respectifs de chaque outil. Ainsi, une note a été attribuée à chacun d'entre eux sur 10 macro-critères résumant ceux abordés en détail dans notre grille de test :

– La société (pérennité de l'éditeur, le pays d'origine…)
– le produit (architecture, coût du produit, prise en main)
– l'accès aux données (volumétrie, pré-formatage)
– les outils linguistiques (présence et la qualité de l'outil linguistique)
– l'automatisation (nécessité d'une intervention manuelle et qualité de l'outil fourni)
– la réduction des dimensions (qualité et diversité des méthodes pour transformer le tableau lexical)
– les méthodes de classification (diversité des méthodes pour classer les documents ou dégager les thèmes abordés)
– les méthodes de classement (diversité des méthodes pour réaliser des modèles de classement automatique)
– la lecture des résultats (présentation et lisibilité des résultats ou aides à l'interprétation)
– et enfin le rapport (présence et qualité du rapport d'étude produit automatiquement par le logiciel)

3.2. *L'interprétation*

Les quatre produits se positionnent très différemment.

Pour Alceste, la partie pré-traitement des données est automatique et efficace (enrichissement important des données grâce aux outils linguistiques). Les classes thématiques obtenues sont homogènes et leur caractérisation avec des variables exogènes, si elles existent, est performante et utile. Le rapport d'analyse, généré automatiquement, est également un atout. Les deux principaux problèmes résident dans la volumétrie limitée et dans l'absence de méthodes de modélisation (aucune méthode de classement par exemple). Alceste est ainsi un outil très efficace pour détecter rapidement les thèmes d'un corpus, mais ne peut se positionner comme un outil de text mining.

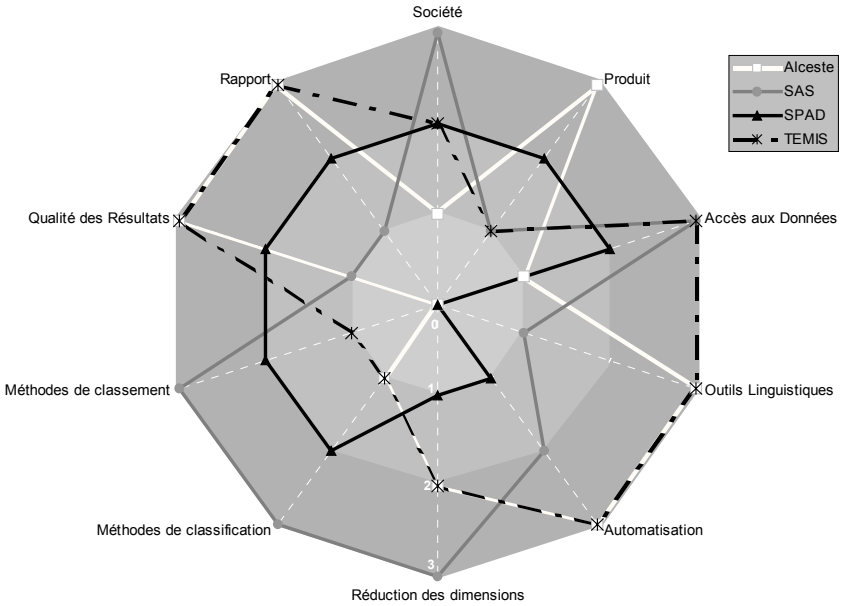


Figure 1. Positionnement des quatre outils de Text Mining

À l'opposé, SAS Text Miner possède un grand nombre de méthodes de modélisation et la volumétrie n'est limitée que par les caractéristiques de la machine. Par contre, pour la partie linguistique, il faut attendre une prochaine version pour véritablement tester ses possibilités (résultats erronés, apparemment non testé pour le français). L'aide à l'interprétation des résultats est également très décevante (pas de caractérisation des classes obtenues, pas de classement par pertinence des documents d'une classe, interface de visualisation peu ergonomique…). Cet outil de text mining s'adresse clairement à des « data miners » confirmés, connaissant déjà le produit de data mining SAS Enterprise Miner dans lequel est inclus le module de text mining.

La suite logicielle TEMIS Insight Discoverer se positionne comme un outil de text mining, permettant une analyse efficace du texte et supportant une grosse volumétrie. Il ne possède cependant pas de fonctionnalités statistiques permettant par exemple de caractériser les classes obtenues par la classification du texte avec des variables illustratives non textuelles. Afin d'utiliser toutes les potentialités du produit, il est pour le moment nécessaire de connaître un langage de programmation permettant de manipuler les données.

Le positionnement de SPAD/CRM se situe entre Alceste et Text Miner. Ce produit dispose d'outils d'analyse de données permettant une exploration fine du corpus et d'outils de modélisation permettant de créer des modèles de classement, sur de grands volumes de données. Par contre, il ne possède aucun outil linguistique, et même si l'interface de filtrage des « mots » est assez conviviale, la préparation des données avant analyse est une étape longue et fastidieuse pour les gros corpus.

4. Conclusion

L'objectif premier de cette étude, à savoir la préconisation d'un logiciel pour l'entreprise, a évolué vers une évaluation comparative des logiciels, dans le sens où la diversité des corpus à traiter ainsi que les objectifs sont tels que les outils sont davantage complémentaires que concurrents.

En ce qui concerne les besoins spécifiques et les corpus d'EDF, Alceste demeure l'outil privilégié de dépouillement et d'exploration des réponses aux questions ouvertes d'enquêtes de satisfaction, d'une volumétrie réduite et présentant des variables explicatives illustratives. La suite logicielle TEMIS Insight Discoverer est quant à elle bien adaptée au traitement des champs commentaires de notre base de données des contacts clients dont la volumétrie dépasse les 100 000 documents et dans lesquels on trouve un langage très spécifique (vocabulaire technique, utilisation d'abréviations...).

Cette évaluation conforte l'idée d'utiliser un protocole de test détaillé, afin d'éprouver un ensemble de fonctionnalités incontournables à la fois en fonction des objectifs visés par l'utilisation d'un outil de text mining, des corpus à analyser mais également du profil de l'utilisateur (langage naturel, statisticien, data miner). De plus, à l'issue de cette dernière, il semble que les outils dont la vocation initiale était l'analyse du texte restent les plus performants.

La grille de test présentée a pour vocation d'évoluer vers un véritable protocole de test d'outils de text mining, d'une part en testant d'autres types d'outils (outils dédiés à la veille par exemple) ou d'autres types de corpus (corpus multilingues) et d'autre part en choisissant un panel d'utilisateurs de niveaux de connaissances variables.

Références

Aït Mokhtar S., Hagège C. et Sàndor A. (2003). Problèmes d'intersubjectivité dans l'évaluation des analyseurs syntaxiques. In *Actes de TALN 2003*. www.sciences.univ-nantes.fr/irin/taln2003/articles/eval1.pdf/

Barthel M.P., Khouas L., Sanford E. et Couillault A. (2002). Évaluation automatique pour résumé automatique. In *Journées d'étude de l'ATALA sur résumés de texte automatiques : solutions et perspectives.* http://www.atala.org/je/021214/Barthel.pdf

Brugidou M., Escoffier C., Folch H., Lahlou S., Le Roux D., Morin-Andréani P. et Piat G. (2000). Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles. In *Actes des JADT 2000.*

CXP. PackExperts 2001 « Business Intelligence : outils de Data Mining », société CXP International. 19-21 rue du rocher, 75008 PARIS.

Nugier S. Garrouste D., Peradotto A. et Quatrain Y. (2003). Grille de test de logiciels de text mining. Note Interne à EDF. Disponible à la demande.

Extending the Cochran rule for the comparison of word frequencies between corpora

Paul Rayson[1], Damon Berridge[2], Brian Francis[2]

[1]Computing Department, Lancaster University, Lancaster, LA1 4YR, UK.
paul@comp.lancs.ac.uk
[2]Centre for Applied Statistics, Lancaster University, LA1 4YF, UK.

Abstract

We first describe a number of inter-related issues that need to be considered by the researcher when comparing frequencies of linguistic features in two or more corpora. We then describe the chi-squared and log-likelihood tests used in previous research for the comparison of word frequencies. Our focus, in this paper, is on the issue of reliability of the statistical tests, and we describe simulation experiments to compare the reliability of the chi-squared and log-likelihood statistics under conditions of different-sized corpora and probability of a word occurring in text. We observe that the Cochran rule provides a good guide to accuracy of both statistics in general, but in some cases it needs to be extended. We conclude by recommending higher cut-off values for the Cochran rule at the 5%, 1% and 0.1% levels. In order to extend applicability of the frequency comparisons to expected values of 1 or more, use of the log-likelihood statistic is preferred over the chi-squared statistic, at the 0.01% level. The trade-off for corpus linguists is that the new critical value is 15.13.

Keywords: word frequency, chi-squared test, log-likelihood test, corpus linguistics.

1. Introduction

In recent years corpus-based techniques have increasingly been used to examine issues in language variation, that is, to compare language usage across corpora, users, genres, etc. Comparison of one-million word corpora is becoming common even for beginners in corpus linguistics, with the increasing availability of corpora and the reasoning that one million words gives sufficient evidence for mid- to high-frequency words. However, with the production of large corpora such as the British National Corpus (BNC) containing one hundred million words (Aston and Burnard, 1998), frequency comparisons are available across several millions of words of text (Leech, Rayson and Wilson, 2001). Sufficient data for the investigation of relatively infrequent phenomena is still problematic. However, research is still continuing on small corpora (Ghadessy, Henry and Roseberry, 2001).

There are two main types of corpus comparison:

- A: comparison of a sample corpus with a large(r) standard corpus (e.g. Scott, 2000)
- B: comparison of two (roughly-) equal sized corpora (e.g. Granger, 1998)

There are also a number of inter-related issues that need to be considered when comparing two (or more) corpora[1]: representativeness, homogeneity within the corpora, comparability of the corpora, and choice and reliability of statistical tests (for different sized corpora and other factors).

[1] Alongside practical issues such as the cost and time taken in obtaining, or collecting such corpora.

In the first type (A), we refer to the large(r) corpus as a 'normative' corpus since it provides a text norm (or general language standard) against which we can compare. These two main types can be extended to the comparison of more than two corpora. For example, we may compare one normative corpus to several smaller corpora at the same time, or compare three or more equal sized corpora with each other. In general, however, this makes the results more difficult to interpret. Biber (1993: 243) states that 'a corpus must be representative in order to be appropriately used as the basis for generalizations concerning a language as a whole'. Representativeness, in this sense, is seen as a particularly important attribute for a normative corpus when comparing a sample corpus to a 'general language' corpus (such as the BNC) that contains sections from many different text types and domains. To be representative of the language as a whole, a corpus should contain samples of all major text types (Leech, 1993) and, if possible, be in some way proportional to their usage in 'every day language' (Clear, 1992). This first type of comparison (A) is intended to discover features in the research corpus which have significantly different usage (i.e. frequency) to that found in 'general' language. Representativeness can also apply to specialised corpora, whenever the researcher wants to make a claim about language use in a particular genre or domain, rather than the language as a whole.

The second type of comparison (B) is one that views corpora as equals. It aims to discover features in the corpora that distinguish one from another. Homogeneity (Stubbs, 1996: 152) within each of the corpora is important here since we may find that the results reflect sections within one of the corpora that are unlike other sections in either of the corpora under consideration (Kilgarriff, 1997). Comparability is of interest too, since the corpora should have been sampled for in the same way; in other words, the same stratified sampling method and with, if possible, randomised methods of sample selection. This is the case with the Brown and LOB corpora (introduced below), since LOB was designed to be comparable to the Brown corpus, and neither corpus was designed to be homogeneous.

The final issue, which we address in this paper, is the one regarding the reliability of the statistical tests in relation to the size of the corpora under consideration. In the next section, we will describe two of the statistical tests previously used, chi-squared and log-likelihood.

2. Background

One of the largest early studies was the comparison of one million words of American English (Brown corpus) with one million words of British English (LOB corpus) by Hofland and Johansson (1982). A statistical goodness-of-fit test, the chi-squared test ($\chi 2$), was used to compare word frequencies across the two corpora. The chi-squared test statistic was calculated as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad \text{where} \quad E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

where Oi is the observed frequency, Ei is the expected frequency, and Ni is the total frequency in corpus i (i in this case takes the values 1 and 2 for the LOB and Brown corpora respectively). Hofland and Johansson marked any resulting chi-squared values that indicated that a statistically significant difference at the 5%, 1%, or 0.1% level had been detected between the frequency of a word in American English and in British English. In some cases the expected frequency in at least one of the two corpora was too low for their calculation. The null hypothesis of the test is that there is no difference between the observed frequencies

of a word in the two corpora. Note that even if the null hypothesis is not rejected, we cannot conclude that it is true. The cut-off value corresponding to the chosen degree of confidence may not be exceeded, but this only indicates there is not enough evidence to reject the null hypothesis. Critical values for the chi-squared statistic are listed in statistical tables such as those in Barnett and Cronin (1986). For example, the critical value for the 5% level, usually shown as 0.05 in the tables, is 3.84 at 1 degree of freedom. Pearson (1904) originally suggested the chi-squared test using the statistic described above for testing the independence of two variables. It is applicable to a general two-dimensional contingency table with r rows and c columns. The number of degrees of freedom (d.f.), which is used when looking up critical values, is the number of independent terms given that the marginal totals in the table are fixed. In corpus linguistics, we usually use a 2×2 table to compare frequencies of words or other linguistic features between two corpora, so d.f. as calculated by (r-1)×(c-1) is equal to 1. In this specific case, the 2×2 contingency table is as shown in Table 1.

Table 1. Contingency table for the chi-squared test

	Corpus one	**Corpus two**	**Total**
Frequency of feature	a	b	a+b
Frequency of feature not occurring	c	d	c+d
TOTAL	a+c	b+d	N=a+b+c+d

Hence, we can calculate the chi-squared statistic (χ^2) as follows:

$$\chi^2 = N(ad-bc)^2/((a+b)(c+d)(a+c)(b+d))$$

Kilgarriff (1996a, 1996b and 2001) points out that in the Brown versus LOB comparison many common words are marked as having significant chi-squared statistics, and that because words are not selected at random in language we will always see a large number of differences in two such text collections. He selects the Mann-Whitney test that uses ranks of frequency data rather than the frequency values themselves to compute the statistic. The drawback of the Mann-Whitney test is that it omits 92% of the types in the joint corpus, as reported in Rayson (2003). Numerous other authors have used the chi-squared test to determine significant frequency differences of individual words or other linguistics features, between two corpora (for example Woods *et al*., 1986: 140; Virtanen, 1997; Oakes, 1998: 26; Roland *et al.*, 2000). It is widely understood that the chi-squared statistic becomes unreliable when the expected frequency is too small, i.e. less than 5 (Butler 1985: 117; Woods *et al.*, 1986: 144), and sometimes the same limit is applied to the observed frequencies (De Cock, 1998; Nelson *et al*., 2002: 277). Cochran (1954) suggested a rule that 4 in 5 (80%) of the expected values in an r×c table should be 5 or more. In the 2×2 table case, this means all cells should have expected values of 5 or more. Everitt (1992: 39) cites other more recent work than Cochran which suggests that this rule is too conservative. Butler (1985: 117) suggests a solution to this is to combine frequencies until the combined classes have an expected frequency of 5 or more, likewise Nelson et al (2002: 277) for the observed frequencies, but Everitt (1992: 41) argues against this practice.

Dunning (1993) reports that we should not rely on the assumption of a normal distribution when performing statistical text analysis and suggests that parametric analysis based on the binomial or multinomial distributions is a better alternative for smaller texts. Hence Dunning

proposes the log-likelihood ratio as an alternative to Pearson's chi-squared test, and he demonstrates this for the extraction of significant bigrams from text. Everitt (1992: 72) also mentions that the chi-squared statistic is "easily shown to be an approximation to" the log-likelihood for large samples. The two statistics take similar values for many tables. Cressie and Read (1984) show that Pearson's χ^2 (chi-squared) and the likelihood ratio G^2 (Dunning's log-likelihood) are in fact two statistics in a continuum defined by the power-divergence family of statistics. Scott (2001) uses the log-likelihood statistic in his keywords procedure. For the 2×2 case (in Table 1), the log-likelihood ratio is calculated as follows:

$$G^2 = 2\ (a\ln a + b\ln b + c\ln c + d\ln d + N\ln N - (a+b)\ln(a+b) - (a+c)\ln(a+c) - (b+d)\ln(b+d) - (c+d)\ln(c+d))$$

Other statistical tests are described in the literature, for example using the Yates' correction to the chi-squared test and Fisher's exact test, and these and other tests are applied in the information retrieval domain. For reasons of space, we omit descriptions of them here, but they are reviewed in more detail in Rayson (2003). In a different application area, that of extracting groups of associated words, Weeber et al (2000) use a combination of the log-likelihood ratio and Fisher's exact test for the full word frequency range. The 2×2 tables for comparing word frequencies across corpora are, in general, not as heavily skewed as for bigram data, so the log-likelihood ratio is suitable for our use. We can now examine situations where the contingency table may become skewed.

The research question we attempt to answer in this paper is as follows. Can we rely on the generally accepted Cochran rule for large variations in corpus size, frequency of words, and ratio of size of the two corpora under comparison? The usual approach for a linguist is to avoid comparisons for low frequency words and in a non-normative test to make the corpora the same size. We wish to test whether we can perform statistically reliable comparisons of corpora of different sizes, and include a larger portion of the frequency profile in our tests.

3. Methodological approach

In order to test the reliability of the chi-squared and the log-likelihood statistics and the prediction ability of the Cochran rule we carried out a large number of simulation experiments (i.e. with simulated contingency tables)[2]. For each experiment, there were three experimental conditions which determined the characteristics of the comparison between the corpora. These were:

1. The ratio of the sizes of the two corpora (R). We assume that one corpus is the normative corpus, and therefore the comparison corpus is less than or equal in size to the normative corpus. Seven different ratio values were used, from 1:1 to 200:1.
2. The probability of the word occurring in text (p). This probability was allowed to vary from 1 in 500 to 1 in 1,000,000 to reflect a good range observed in the BNC. To give context to these values, the probability of 'the' occurring in standard British English is 1/16, the probability of 'and' is 1/37 and the probability of 'reliable' is 1/44823 (estimated from the BNC). Ten different probability values were used. We assumed that the probability of the target word was the same in the normative text and in the com-

[2] The GLIM4 statistical software package was used to do this (Francis, Green and Payne, 1993) but the simulations can be coded in any programming language with access to a reliable random-number generator.

parative text, as we wished to test the distribution of the test statistics χ^2 and G^2 under the null hypothesis of no difference in probabilities.

3. The size of the normative corpus (C). This was allowed to vary from 100,000 to 100 million words, taking eight different values. The BNC is at the top end of this range, with 100 million words.

We assumed that the probability of the target word was the same in the normative text and in the comparative text. For each experiment, C, R, and p were all fixed, and 10,000 simulated 2×2 tables of word frequencies were generated. Table 2 below shows the expected values of the 2×2 table. For each simulated table, the chi-squared test statistic χ^2 and the log-likelihood test statistic G^2 were calculated, and at the end of each experiment the χ^2 statistics were ranked in ascending order; the G^2 statistics were similarly ranked. The 95th, 99th, 99.9th and 99.99th percentiles were then determined. The experiment was replicated one hundred times to obtain empirical estimates of the standard deviation of the percentiles.

Table 2. Expected values of the 2×2 tables considered in the experiment

	Normative corpus	Comparative corpus	Total
No. of target words	pC	pRC	p(1+R)C
Number of non-target words	(1-p)C	(1-p)RC	(1-p)(1+R)C
Total number of words	C	RC	(1+R)C

Finally, for each experiment, we determined whether the usual Cochran rule would determine if the simulated values were unreliable. As the rule is based on expected values, and not actual simulated values, this could be determined once at the beginning of each experiment. From Table 2, the smallest expected value is pRC (as $P < 1$ and $R \leq 1$ by design). Thus, if pRC was less than 5, the Cochran rule was held to be true. A total of 560 (7×10×8) experiments were carried out, covering each possible combination of the three experimental factors. With one hundred replicates of each experiment, the entire study generated 560,000,000 (10,000×100×560) simulated 2×2 tables.

4. Results and discussion

Table 3 shows some typical output from the simulation experiments, giving the results for the 95th percentile of the χ^2 test statistic where p=1/16000. For each combination of normative corpus size C and ratio R, the table entries give the mean and standard deviation (in brackets) of the 100 simulated values of the test statistic. We define the test statistic to be accurately represented by the chi-squared distribution if the 95th percentile of the chi-squared distribution on 1 degree of freedom (3.840) is contained in the interval defined by the mean plus or minus two standard deviations. Cells of the table in bold font show where this condition does not hold. It can immediately be observed that the test statistic is accurate in most of the table, but not in the top left hand corner. This corner is characterised by small expected values in the 2×2 table, and examination of Table 4, which gives the smallest expected value, indicates that the Cochran condition provides a good guide to accuracy in this case. Cells of the table in bold font show where the Cochran rule is true.

Table 3. Means and standard deviations of the 95th simulated percentile of the chi-squared test statistic under independence for various 2×2 tables with p=1/16000[3]

p=1/16000	**Ratio R**						
Corpus size C	**1 / 200**	**1 / 100**	**1 / 20**	**1 / 10**	**1 / 5**	**1 / 2**	**1 / 1**
100,000	**0.060**	**8.772**	3.718	4.068	**3.410**	**3.528**	3.752
	(0.000)	**(0.664)**	(0.206)	(0.192)	**(0.098)**	**(0.069)**	(0.062)
500,000	**5.0945**	**2.615**	**3.400**	**3.605**	3.763	3.836	3.831
	(0.070)	**(0.100)**	**(0.119)**	**(0.037)**	(0.074)	(0.081)	(0.070)
1,000,000	**2.253**	**3.505**	**3.523**	3.736	3.805	3.828	3.823
	(0.069)	**(0.064)**	**(0.055)**	(0.059)	(0.070)	(0.081)	(0.074)
5,000,000	**3.707**	**3.324**	3.801	3.811	3.831	3.850	3.845
	(0.048)	**(0.034)**	(0.075)	(0.075)	(0.068)	(0.076)	(0.084)
10,000,000	**3.244**	3.739	3.813	3.834	3.840	3.842	3.837
	(0.019)	(0.101)	(0.067)	(0.065)	(0.079)	(0.065)	(0.082)
20,000,000	3.720	3.752	3.813	3.842	3.851	3.847	3.854
	(0.073)	(0.077)	(0.073)	(0.076)	(0.070)	(0.070)	(0.078)
50,000,000	3.773	3.809	3.829	3.833	3.849	3.836	3.845
	(0.047)	(0.071)	(0.075)	(0.072)	(0.081)	(0.074)	(0.080)
100,000,000	3.820	3.826	3.835	3.843	3.844	3.845	3.831
	(0.068)	(0.072)	(0.071)	(0.067)	(0.074)	(0.071)	(0.077)

Table 4. Smallest expected values in the 2×2 tables when p=1/16000

p=1/16000							
Corpus size	**Ratio R**						
C	**1 / 200**	**1 / 100**	**1 / 20**	**1 / 10**	**1 / 5**	**1 / 2**	**1 / 1**
100,000	**0.031**	**0.063**	**0.313**	**0.625**	**1.250**	**3.125**	6.250
500,000	**0.156**	**0.313**	**1.563**	**3.125**	6.250	15.625	31.250
1,000,000	**0.313**	**0.625**	**3.125**	6.250	12.500	31.250	62.500
5,000,000	**1.563**	**3.125**	15.625	31.250	62.500	156.250	312.500
10,000,000	**3.125**	6.250	31.250	62.500	125.000	312.500	625.000
20,000,000	6.250	12.500	62.500	125.000	250.000	625.000	1250.000
50,000,000	15.625	31.250	156.250	312.500	625.000	1562.500	3125.000
100,000,000	31.250	62.500	312.500	625.000	1250.000	3125.000	6250.000

[3] Standard deviations are in parentheses. Cells where the 95% critical value of the chi-squared distribution on 1 degree of freedom (3.84) lies outside interval defined by the mean plus or minus two standard deviations are defined as inaccurate and shown in bold.

Of course, Table 3 is a small portion of the output generated. Similar tables were generated for G^2 and χ^2; for the 99th, 99.9th and 99.99th percentiles as well as the 95th percentile, and for each of the ten values of the proportion p. Table 5 summarises the results into a single display. The table consists of an array of ten rows and five columns. The first four columns show the accuracy of the tests at the 5%, 1%, 0.1% and 0.01% significance level, and the final column the standard Cochran rule. Each of the ten rows represents a different proportion. Within a cell of the array, an 8×7 grid shows the simulated accuracy. Rows of each grid represent the eight corpus sizes in ascending order, and the columns of the grid represent the seven ratios, again in ascending order. The symbols used are: ■ both tests inaccurate; ◘ chi-squared test inaccurate; ▣ likelihood ratio test inaccurate; □ both tests accurate. A ✖ indicates that the smallest expected value of the generated table is less than 5. The critical values we used were 3.84 (5% level), 6.63 (1% level), 10.83 (0.1% level) and 15.13 (0.01% level which is not usually listed in published tables).

It can be seen from Table 5 that the overall pattern in each 8×7 grid is similar to that in Table 3. The chi-squared and likelihood ratio test statistics are accurate in much of each grid apart from the top left hand corners. There is also an observable trend as we move down the ten rows from the top to the bottom of the overall table. The trend moving down the overall table is for fewer cells in each 8×7 grid to be marked indicating fewer inaccurate tests as the proportion increases. A trend moving from left to right (from 5%, 1% to 0.1% and 0.01%) in the overall table is less easy to detect. For all the proportions, there are slightly fewer inaccurate tests observed at the 99.99th percentile (0.01%) than at the 95th percentile (5%) level.

Let us now contrast the accuracy of the two tests. When one test rather than both is inaccurate, and this occurs in 246 grid cells, the likelihood ratio test is inaccurate 81 times, and the chi-squared test is inaccurate 165 times. There is an overall trend from left to right which distinguishes the two tests. It is most clear at the 0.01% level in which, with the exception of 25 cells, the likelihood ratio is accurate in all the cells. This compares with 120 cells where the chi-squared test is inaccurate. Examining the data underlying the results at the 0.01% level in particular, the simulations for the likelihood ratio showed much smaller standard deviations than for the chi-squared test, as well as the mean statistic for the chi-squared test being skewed. The likelihood ratio test remains accurate for very heavily skewed tables at the 0.01% level, even for example, in one of the cells in the first grid at the 0.01% level represents a simulation of a word occurring once in a 1,000,000 word corpus as compared to the word occurring 100 times in a 100 million word corpus. The cells are the top right of the grids at the 0.01% level are false positives since the simulated critical values fall below the listed one.

If for each row we compare the four grids showing the accuracy of the tests to the grid showing the Cochran rule, we can see that the Cochran rule provides a good guide to accuracy in most cases. However, there are some cases (marked with a shaded background) which show inaccurate tests that are not covered by the Cochran rule. These are cases where the smallest expected value in the generated table is 5 or greater. The cells in question show inaccuracy in both chi-squared and likelihood ratio tests. At the 5% level the largest expected value showing an inaccurate test is 12.5. At the 1% level we observe a value of 10, at the 0.1% level it is 7.8125 and at the 0.01% level the largest expected value coinciding with an inaccurate test is 6.25. This suggests that the Cochran rule needs to be extended.

The statistics have more evidence from larger corpora and can therefore detect smaller differences in frequency. It is a feature of the tests that values are greater in larger corpora. So, for example, a χ^2 value of 500 obtained from a 1 million versus 9 million word normative test is

not comparable with the same χ^2 value of 500 obtained from a 10,000 versus 90,000 word comparison. This does not mean the tests are flawed, but that we must be careful when considering their results.

5. Conclusion

There are several conclusions we can draw from our experiments. The statistical tests are accurate for the most part with various combinations of corpus size, word probability and ratio of corpora. From the point of view of both of the statistical tests there are no problems comparing unbalanced sized corpora as long as we avoid low expected values in the contingency table. At the 5% level, the Cochran rule should be extended to ensure expected values are 13 or more. At the 1% level, the Cochran rule should be extended to ensure expected values are 11 or more. At the 0.1% level, the Cochran rule should be extended to ensure expected values are 8 or more. The usual Cochran rule is sufficient at the 0.01% level for the chi-squared test. However (ignoring false positives), we can safely lower the Cochran rule at the 0.01% level for the log-likelihood test to expected values of 1 or more. The trade-off is that the critical value is higher than at the usual 5% level at 15.13.

There is a difference between establishing statistical significance and practical significance. In carrying out tests of significance we should always bear in mind the other issues relevant to corpus comparison as listed in the introduction which are equally important as determining significance: representativeness, homogeneity and comparability. The final issue, that of choice and reliability of statistical tests, has been addressed directly in this study. In any comparison experiments one should keep all these issues in mind when interpreting the results. For example, if we chose to compare a written corpus with a spoken corpus, it is very likely that lexical and grammatical differences between the spoken and written language will be exposed as well as differences in domain or content that we may wish to focus on.

References

Aston G. and Burnard L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press.

Barnett S. and Cronin T. M. (1986). *Mathematical formulae for engineering and science students, fourth edition*. Longman.

Biber D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, vol. (8/4), Oxford University Press: 243-257.

Butler C. (1985). *Statistics in linguistics*. Blackwell.

Clear J. (1992). Corpus sampling. In Leitner G. (Ed.) *New directions in English language corpora*. Mouton-de-Gruyter: 21-31.

Cochran W. G. (1954). Some methods for strengthening the common $\chi 2$ tests. *Biometrics*, vol. (10): 417-451.

Cressie N. and Read T.R.C. (1984) Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. (46/3): 440-464.

De Cock S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, vol. (3/1), John Benjamins: 59-80.

Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol. (19/1): 61-74.

Everitt B.S. (1992). *The analysis of contingency tables, 2nd edition*. Chapman and Hall.

Francis B., Green M. and Payne C. (1993) *The GLIM4 system*. Oxford University Press.

Ghadessy M., Henry A. and Roseberry R.L. (Eds) (2001). *Small corpus studies and ELT: theory and practice*, John Benjamins.

Granger S. (1998). The computer learner corpus: a versatile new source of data for SLA research. In Granger S. (Ed.), *Learner English on Computer*. Longman: 3-18.

Hofland K. and Johansson S. (1982). *Word frequencies in British and American English*. The Norwegian Computing Centre for the Humanities.

Kilgarriff A. (1996a). Which words are particularly characteristic of a text? A survey of statistical approaches. In Evett L.J. and Rose T.G. (Eds), *Language Engineering for Document Analysis and Recognition (LEDAR), AISB96 Workshop proceedings, Brighton, England*. Faculty of Engineering and Computing, Nottingham Trent University: 33-40.

Kilgarriff A. (1996b) Why chi-square doesn't work, and an improved LOB-Brown comparison. *ALLC-ACH Conference*.

Kilgarriff A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings 5th ACL workshop on very large corpora*. Beijing and Hong Kong.

Leech G. (1993). 100 million words of English: a description of the background, nature and prospects of the British National Corpus project. *English Today* 33, Vol. (9/1). Cambridge University Press.

Leech G., Rayson P., and Wilson A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. Longman.

Nelson G., Wallis S. and Aarts B. (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English*. John Benjamins.

Oakes M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.

Pearson K. (1904). On the theory of contingency and its relation to association and normal correlation. *Biometric Series*, vol. (1). Drapers' Co. Memoirs.

Rayson P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Roland D., Jurafsky D., Menn L., Gahl S., Elder E. and Riddoch C. (2000). Verb Subcategorization Frequency Differences between Business-News and Balanced Corpora: the role of verb sense. In *proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000).* 1-8 October 2000, Hong Kong: 28-34.

Scott M. (2000). Focusing on the text and its key words. In Burnard L. and McEnery T. (Eds) *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora*. Peter Lang: 104-121.

Scott M. (2001). Mapping key words to problem and solution. In Scott M. and Thompson G. (Eds) *Patterns of Text: in honour of Michael Hoey*, Benjamins: 109-127.

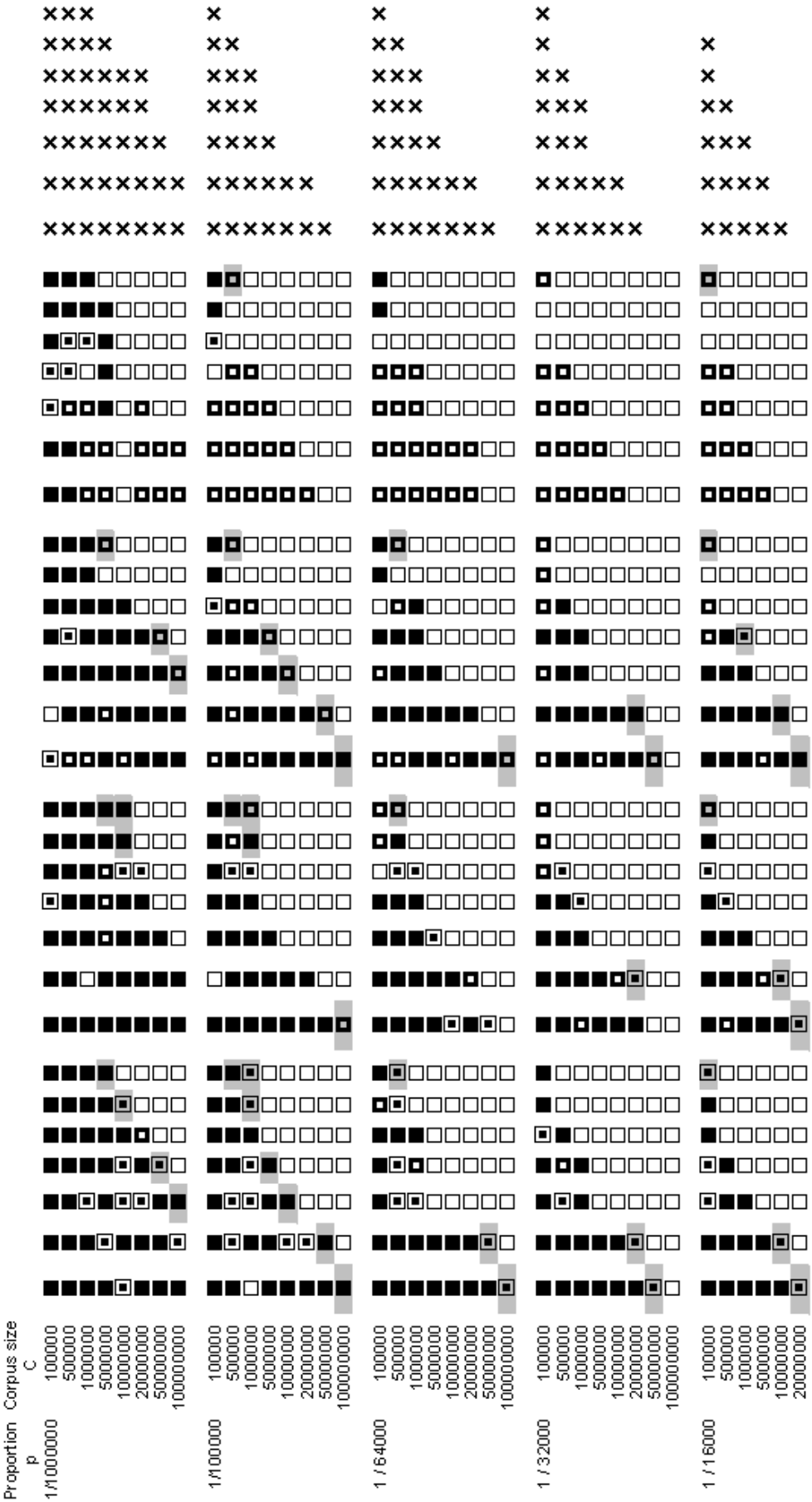
Stubbs M. (1996). *Text and corpus analysis: computer-assisted studies of language and culture*. Blackwell.

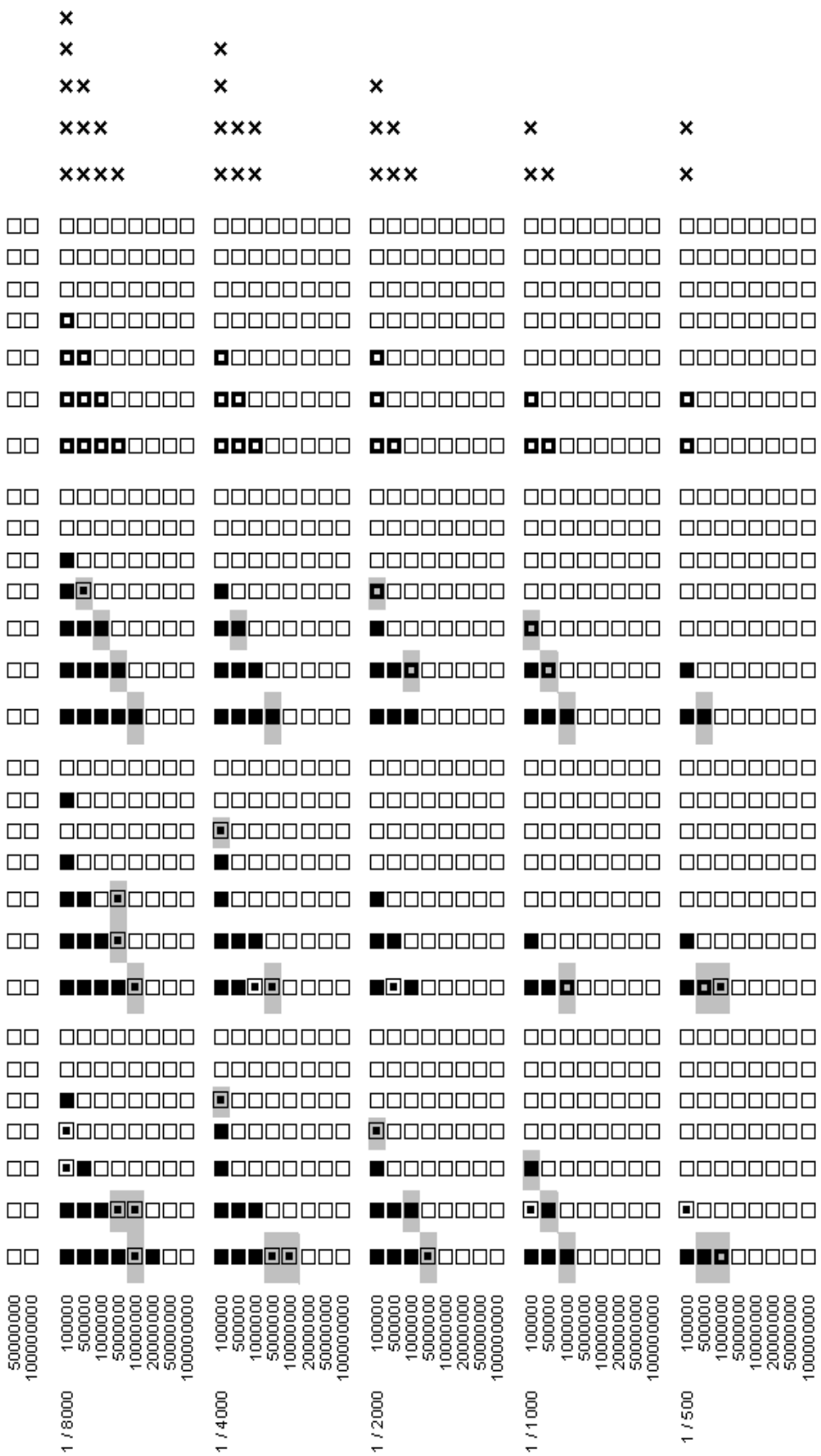
Virtanen T. (1997). The progressive in NS and NNS student compositions: evidence from the International Corpus of Learner English. In Ljung M. (Ed.) *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17)*, Stockholm, May 15-19, 1996. Rodopi: 299-309.

Weeber M, Vos R. and Baayen R.H. (2000). Extracting the Lowest Frequency Words: Pitfalls and Possibilities. *Computational Linguistics*, vol. (26/3). MIT Press: 301-317.

Woods A., Fletcher P. and Hughes A. (1986). *Statistics in language studies*. Cambridge University Press.

Table 5 Accuracy of chi-squared and likelihood ratio tests at the 5%, 1%, 0.1% and 0.01% level.




1 / 8000
1 / 4000
1 / 2000
1 / 1000
1 / 500
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000

Diversity of Vocabulary and Homogeneity of Style in *Tirant lo Blanc*

Alex Riba, Josep Ginebra

Dep. d'Estadística, U.P.C. – Diagonal 647 – 08028 Barcelona – España.
{alex.riba, josep.ginebra}@upc.es

Abstract

Tirant lo Blanc is the main work in catalan literature, and it is considered by some to be the first modern novel in Europe. To help settle the debate around its authorship, we analyze the evolution of the diversity of various measures of the vocabulary used, evaluated first on consecutive blocks of text of 1000 words and second on the chapters of the book. The analysis reveals a sharp boundary, that is estimated through change-point techniques to be around chapter 382, and might indicate a switch in authorship. The language used after that point is a lot less rich and diverse than before.

Keywords: stylometry, diversity indexes, Simpson index, change point estimation, switch regression model.

1. Introduction

Tirant lo Blanc is a chivalry book hailed to be "the best book of its kind in the world" by Cervantes in Don Quixote, and considered by some to be the first modern novel in Europe, (see Vargas Llosa, 1991 and 1993). It is the main work in catalan literature, and its main body was written between 1460 and 1464, but it was not printed until 1490. There has been an intense and long lasting debate around its authorship originating from conflicting information given in its first edition. Where in its dedicatory letter it is stated that the whole book is the work of Martorell (1413-1468), in the colophon of that first edition it is stated that the last one fourth of the book was written by Martí Joan de Galba, after the death of Martorell.

The degree of participation of Galba in the book has been the object of a long lasting and intense debate. Articles like Martínez Martínez (1916), Entwistle (1927), Moll (1933), Menéndez Pelayo (1943), Riquer (1947), Coromines (1956), Nicolau d'Olwer (1961), Goertz (1967), Bosch (1987), Ferrando (1989 and 1995), Rubiera (1990 and 1992) and Wittlin (1990 and 1993) favor a literal interpretation of the "colophon" defending the hypothesis of a change of author between chapters 300 and 350, and thus that Galba is the author of the last one fifth of the book, composed by 487 chapters. On the other hand, articles like Givanel (1916), Vaeth (1918), Marinesco (1979), Riquer (1990), Hauf (1993), Chiner (1991 and 1993), Badia (1993), Casanova (1994) favor the hypothesis that the whole book was written by a single author, that is, by Martorell. See also Guia (1995 and 1996) for an heterodox interpretation of this question. For a nice overview of the debate around this authorship problem, see Riquer (1990) and Chiner (1993).

Neither of the two candidate authors left any text comparable to the one under study, and therefore one can not use discriminant analysis to help classify the chapters in the book by author. By analyzing the style in sample texts encompassing about ten percent of the book, Ginebra and Cabos (1998) detected heterogeneities that might indicate the existence of two authors. Through

the analysis of the multinomial sequences associated to the distribution of word length and to the use of the most frequent words, Riba (2002), Riba and Ginebra (2003) and Giron, Ginebra and Riba (2004) find a sharp stylistic boundary somewhere between chapters 371 and 382.

In this paper we characterize the style of *Tirant lo Blanc* through the diversity of its vocabulary, measured from the list of all the words that appear in the book, and the number of occurrences of each word. Section 2 presents measures of the diversity of the vocabulary and Section 3 explores the evolution of these measures evaluated in consecutive blocks of text of 1000 words, as well as on the chapters that have very unequal lengths. That reveals a sharp boundary, that is estimated through change-point techniques to be around chapter 382.

2. The Measure of the Diversity of the Vocabulary

To characterize the style of an author through the diversity of its vocabulary, the basic assumption is that he has available a certain stock of words that he uses in his writings, some of which he favors more than others. For each word i, let π_i be the proportion of times that it would be found in a text of infinite length of that author. It is assumed that both the list of words that form the author's hypothetical vocabulary as well as the values of π_i are constant along all the writings under study of an author. This hypothesis will rarely hold whenever these writings embrace different epochs and/or genres.

If we sample texts produced by a writer, his vocabulary will be reflected in the sample frequency of each word that appears in the texts. Let N denote the total number of words in a given text, let n_i denote the number of occurrences of word i, and let $\hat{\pi}_i = n_i/N$ denote the proportion of occurrences of i in that sample text. Obviously N is equal to the sum of n_i over all words i in his vocabulary. If the hypothesis of stationarity applies, the expected value of $\hat{\pi}_i$ is π_i, and therefore it does not depend on N.

Furthermore, given a sample text of length N, let V denote the number of different words and let V_r denote the number of different words appearing exactly r times in the text. Note that the sum of V_r over all integers r is equal to V and that the sum of rV_r over them is equal to N. The proportion of different words appearing r times in a sample text (of length N) is $\hat{p}_r = V_r/V$, and its expected value is denoted by $p_r(N)$. Note that the distribution of $\hat{p}_r$, and in particular its expected value, depends on N. Modelling the vocabulary through $\hat{p}_r$ is more convenient than doing it through $\hat{\pi}_i$, because the support of the random variable $\hat{\pi}_i$ encompasses both the observed vocabulary as well as the unobserved one, and therefore it is unknown, while the support of $\hat{p}_r$ is a subset of the integers.

Given a text with a total of N words, the larger V, the most rich and diverse its vocabulary, and a very simple measure of diversity is V itself. Observe though that the distribution of V depends on N, and in particular note that while N could increase without limit, V is limited by the number words in the vocabulary of the author. Thus, V by itself can only be used to compare texts of the same length.

Given N and V fixed, the larger p_r for small r and the smaller p_r for large r, more diverse its vocabulary. Any diversity measure should respect these two ordering levels. Two alternative diversity measures are V_1, that is the number of words appearing once, and V_2, the number of words appearing twice. Given N, the larger V_1 and V_2, the more diverse the vocabulary, but since their distributions also depend heavily on N (see Sichel, 1986a), they are not convenient when comparing texts of different lengths.

An alternative measure is the entropy of the the vocabulary of a sample text, defined as:

$$H = -\sum_i \hat{\pi}_i log\hat{\pi}_i = -\sum_i \frac{n_i}{N} log\frac{n_i}{N} = -\sum_r \frac{rV_r}{N} log\frac{r}{N}.$$

The larger H, the more diverse the vocabulary, but for a given author, as N increases $E(H)$ decreases and is different from the entropy of the vocabulary of the author. That means that H is not useful either, when comparing texts of different lengths.

Instead, Simpson (1949) proposes measuring diversity through the probability that picking up two words at random from a sample text, they are the same word, D, that can be computed as:

$$D = \frac{\sum_r r(r-1)V_r}{N(N-1)} = \frac{\sum_i n_i(n_i-1)}{N(N-1)}.$$

This index is also called Gini's Index, because it was considered in another context by Gini (1912). The more diverse the vocabulary, the smaller the probability of repeating words, and therefore the smaller D tends to be. What makes the Simpson index, D, more useful than V, V_1 or V_2 is that its expected value depends on the diversity of the vocabulary of the author, but it does not depend on N, and that makes it a very nice tool when comparing texts of different lengths. In fact, the expected value of D is the probability that two words picked up at random from a text of infinite length are the same word,

$$E(D) = \sum_i \pi_i^2.$$

Sichel (1986a) recommends evaluating all these diversity measures based only on words of a special kind, like nouns, prepositions or conjunctions, but we do it based on all words instead. Following the example of Mosteller and Wallace (1984), we do not lemmatize *Tirant lo Blanc* and thus we do not reduce all verbs to their infinitive, all nouns to their singular and all adjectives to the same genre. For more details on the measure of diversity, see Good (1953), Good and Toulmin (1956), Margalef (1958), Efron and Thistedt (1979 and 1987), Patil and Taillie (1982), Pielou (1982), Harris (1982), Rao (1982), Holmes (1985) and Sichel (1986a and 1986b), Manly (1994) and Lebart *et al.* (1998), Alvarez, Bécue and Lanero (2000), Hubert *et al.* (2002).

3. Diversity of the Vocabulary in *Tirant lo Blanc*

Here we use the modern edition of *Tirant lo Blanc* by Riquer (1983), because the original manuscript is not available in digital format. Given the length of this manuscript and the unavailability of a printed reproduction of the first edition of the book, it was beyond our means introduce that edition available in digital format. Even though it would be clearly preferible to do the analysis based on the edition of 1490, we believe that finding a stylistic boundary on a modern edition where some experts believe there is a change of author, is very much worth reporting.

After excluding the titles of the chapters and words in italics, that correspond to quotations in latin, the book amounts to a total of 396349 words. In that edition, the book is broken into 487 chapters of very unequal lengths, with 19 chapters exclusively in italics and a longest chapter of 6521 words.

3.1. *Change-point Analysis through Blocks of Text of Constant Length*

To avoid the effect of the dependency of the distribution of the diversity measures on text length, we divide *Tirant* in 396 blocks of $N = 1000$ consecutive words, discarding the last 349 words of the book from consideration. Some blocks include more that one chapter while some chapters are split in several blocks.

For each of the blocks, we compute the value of V, V_1, V_2, H, and D. Clearly, the larger V and H and the smaller D, the more diverse and rich the vocabulary in the corresponding block. Figure 1 presents the relation between these 396 values of V, V_1, V_2, H, and D. Given that V_2 is not correlated with the other diversity measures, we will not use it in the rest of the analysis.

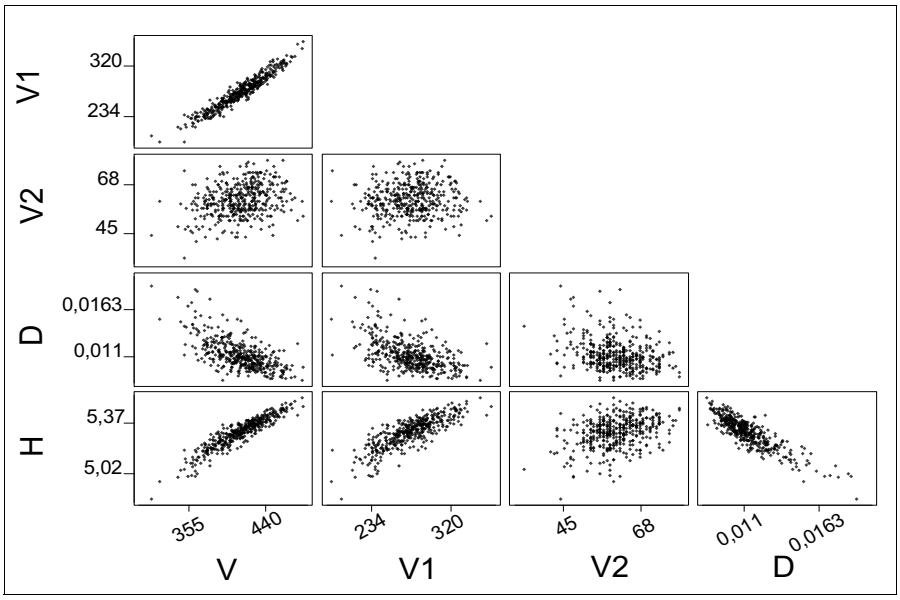


Figure 1. Relation between the values of V, V_1, V_2, H, and D for the 396 blocks of 1000 *consecutive words that form* Tirant lo Blanc.

Figure 1 presents the sequence of values taken by V, H and D in these blocks of constant length. The three indices present a clear shift in their level at around the same point. In the first part of the book, V and H tend to be larger and D tends to be smaller than in the last part of it, and thus the vocabulary of the main part of the book is richer and more diverse than the vocabulary used at the end of the book. Next, the change-point in these sequences of diversity measures is estimated using the maximum likelihood criteria.

Let $y_1, y_2, \ldots, y_n$ be an ordered sequence of mutually independent random variables, with distribution function $F_{\theta_0}(y)$ for $i = 1, \ldots, r$, and distribution function $F_{\theta_1}(y)$ for $i = r+1, \ldots, n$, where θ_0, θ_1 and r are unknown. The estimation of change-point r has been extensively studied for various univariate distributions and very different points of view. For a likelihood based approach, see for example Hinkley (1970 and 71), for a non-parametric approach see Wolfe and Schechtman (1984) while for presentations of the Bayesian approach see Smith (1975 and 80). For a review, see Pettit (1989).

In our case, y_i would be the values V_i, H_i and D_i taken by V, H and D in each block. If all the chapters were written by the same author, it would be reasonable to expect the distributions for V_i, H_i and D_i to be the same for all i. On the the other hand, if there was a change of author in block $i = r$, one might detect a change in distribution at that point. We assume that

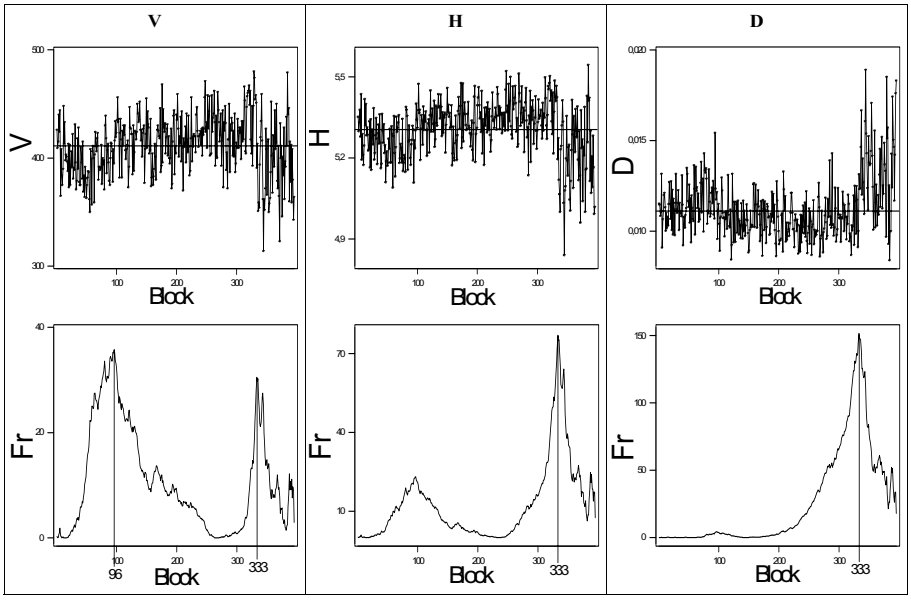


Figure 2. Sequence of the values taken by V, H and D, for the blocks of 1000 *consecutive words. The corresponding values of the F statistic for the single-switch normal regression models, (1), as a function of the switch-point, r, are presented below. Block* 333 *corresponds to chapter* 382.

these observations are conditionally independent and normally distributed, and that the change in distribution translates into a sudden shift in its expected value. In order to estimate r, we fit all the single-switch normal regression models:

$$V_i \sim N(\beta_0^r + \beta_1^r Ir_i, \sigma^2), \tag{1}$$

with $Ir_i = 0$ for $i \leq r$ and $Ir_i = 1$ for $i > r$, that assume $E(V_i) = \beta_0^r$ for $i \leq r$ and $E(V_i) = \beta_0^r + \beta_1^r$ for $i > r$. The change-point r is then estimated to be the one under which (1) fits the observed sequence V_i. As the goodness of fit criteria we use the F-statistic from the ANOVA table, F_r; and the value of r that maximizes F_r is the maximum likelihood estimate of the change-point of a sequence of mutually independent normal observations. We also repeat this exercise for the sequences of H_i and of D_i.

Figure 2 presents the values of F_r as a function of r for these sequences. One can observe how for the three sequences there is a clear maximum for the block numbered $i = 333$, that corresponds to chapter 382. Given that our diversity indices are all positive, we repeated this analysis using single-switch gamma regression models instead of normal ones, and we obtained very similar results to the ones reported here. For a cluster analysis of these data, that reinforces the evidence in favor of the existence of this stylistic boundary, see Riba (2002).

3.2. *Change-point Analysis of the Diversity Measured in Chapters*

When one uses the chapters of the book as the unit of interest instead of blocks of constant length, one needs to take the length of each chapter i, N_i, into consideration.

Figure 3 presents the relation between D_i and N_i and the time plot of D_i for the 425 chapters of more than 200 words. Given that $E(D_i)$ does not depend on N_i, one can estimate the change-point in that sequence using an approach analogous to the one in Section 3. That is, one fits the single-switch linear normal model (1) for each $r = 1, \ldots, 424$, through weighted least squares, (because the variance of D_i is proportional to $1/N_i$), and one estimates the change-point to be

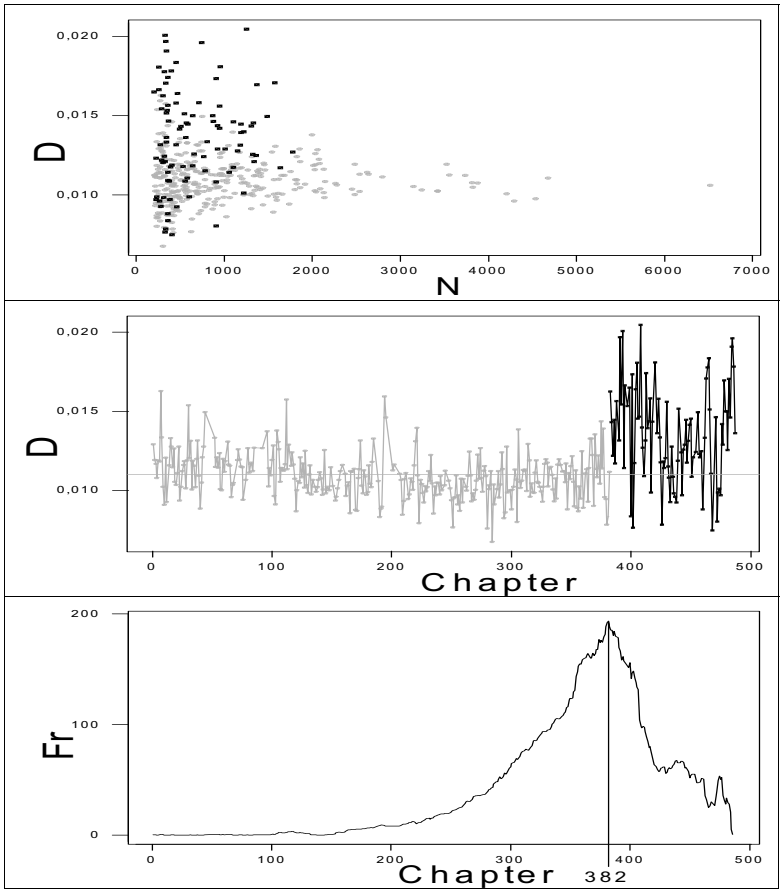


Figure 3. The top panel presents the relation between the Simpson index for chapter i, D_i, and the total number of words in it, N_i. The middle panel presents the sequence of D_i. The bottom panel presents the F statistic for (1) *as a function of r for this sequence. Gray indicates chapters 1 to 382 while black indicates chapters 383 to 487.*

the one that corresponds to the model with the largest F. It turns that the best fit for the observed sequence is again obtained for $r = 382$.

Figure 4 relates the values taken by V and H in chapter i, V_i and H_i, with N_i, thus illustrating the vocabulary growth when the text length grows. From that figure, it is clear that in the logarithmic scales there is a strong linear relationship between V and N and between H and N, even though the vocabulary growth seems to be different for the chapters at the end of the book, in black, than for the ones at the beginning of it, in gray. That relationship is even more linear when using $H_s = H/logN$ against $logN$ instead of $logH$ against $logN$ (see Holmes, 1985).

Here we use the relation between V_i and N_i and between H_i and N_i to characterize the vocabulary of an author,through their vocabulary growth. By searching for the chapter where these relations most likely change, we find new estimates of the stylistic boundary. More specifically, we fit the single-switch linear normal regression model:

$$\log V_i \sim N(\beta_0^r + \beta_1^r \log N_i + \beta_2^r Ir_i + \beta_3^r Ir_i * \log N_i, \sigma^2), \tag{2}$$

for all $r = 1, \ldots, 424$, that assume that $E(\log V_i) = \beta_0^r + \beta_1^r \log N_i$ for $i \leq r$ and $E(\log V_i) = \beta_0^r + \beta_2^r + (\beta_1^r + \beta_3^r) \log N_i$ otherwise. Again, r is estimated to be the change-point for which the corresponding model fits the observed V_i best. We repeat this same estimation procedure for H_i. In both instances, the change-point is estimated to be again $r = 382$.

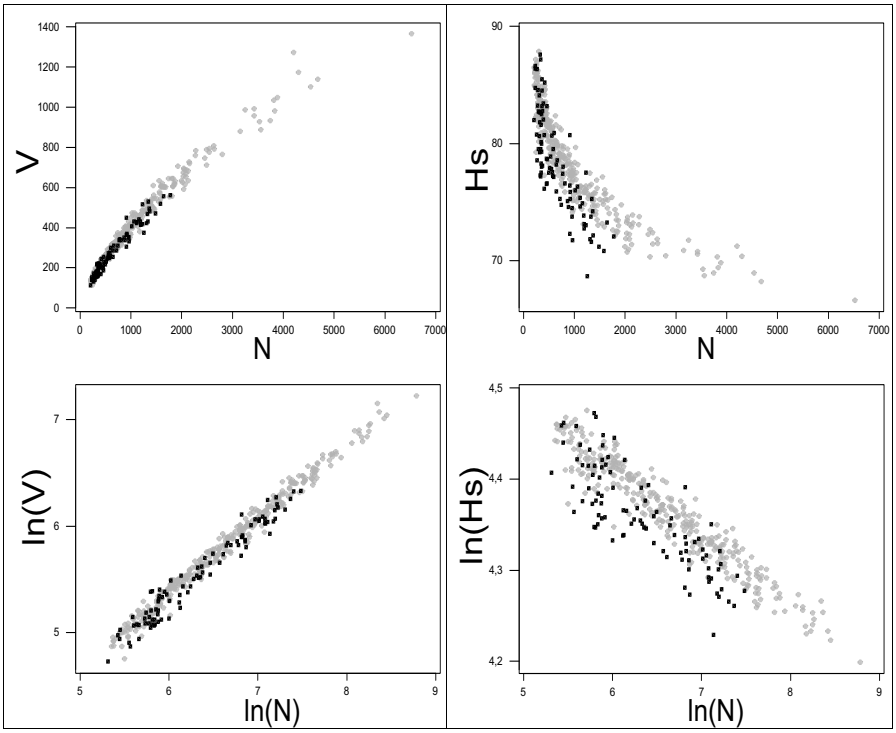


Figure 4. Relation between the number of different words in chapter i, V_i, and N_i, and of H_i and N_i, for the 425 *chapters of more than* 200 *words, illustrating the vocabulary growth in* Tirant lo Blanc. *Gray dots correspond to the chapters before chapter 382, while the black dots identify the ones after it.*

4. Conclusions

Despite the use of a modern edition of *Tirant lo Blanc*, we detect a very clear stylistic boundary near chapter 382, in very close agreement with the result of the statistical analysis on other variables, reported in Riba (2002), Riba and Ginebra (2003) and Giron, Ginebra and Riba (2004). That boundary is placed very close to where is anticipated by the colophon of the first edition of the book, and where it was guessed by some experts on the book. It is important to remark that even though the statistical analysis supports the existence of two authors, it is not up to us to exclude the possibility that the stylistic boundary found could be explained otherwise. Also, the fact that a few chapters at the end of the book have diversity indices that are similar to the ones in the chapters at the beginning of it, might mean that the second author finished the book by filling in material at the end of it.

References

Alvarez R., Bécue M. and Lanero J.J. (2000). Vocabulary Diversity and its Variability: a Tool for the Analysis of Discoursive Strategies – Application to the Investiture Speeches of the Spanish Democraty. In *Actes des JADT 2000 on line at* http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/tocJADT2000.htm

Badia L. (1993). El Tirant en la tardor medieval catalana. In *Actes del Sismposion Tirant lo Blanc*: 35-99.

Bosch R. (1987). Dos llenguatges i dos autors del Tirant lo Blanc. *Daina*, vol (2): 23-34.

Casanova E. (1994). Llengua popular i llengua col·loquial al Tirant lo Blanch. In *Miscel·lània Joan Fuster*, vol. (VIII).

Chiner J. (1991). Batalla a ultrança per Joanot Martorell, *A sol post*, vol (2): 83-127.

Chiner J. (1993). *El viure novel·lesc. Biografia de Joanot Martorell*: 160-166.

Coromines J. (1956). Sobre l'estil i manera de Martí Joan de Galba i els de Joanot Martorell. In *Lleures i converses d'un filòleg*. Club editor (1971): 363-378.

Efron B. and Thistedt R. (1979). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, vol. (63): 435-447.

Entwistle W.J. (1927). Observacions sobre la dedicatòria i primera part del Tirant lo Blanc. *Revista de Catalunya*, vol. (7): 381-398.

Ferrando A. (1989). Entorn de la llengua del Tirant lo Blanc. *Saó*: 24-26.

Ferrando A. (1995). Del Tiran de 1460-1464 al Tirant de 1490. In Paredes J., Nogueras V. and Sánchez L. (Eds), *Estudios sobre el Tirant lo Blanc*: 75-109.

Ginebra J. and Cabos S. (1998). Anàlisi estadística de l'estil literari; Aproximació a l'autoria del Tirant lo Blanc. *Afers*, vol. (29): 185-206.

Gini C. (1912). Variabilitá e Mutabilitá. In *Studi Economico-Giuridici della facoltá di Giurisprudenza dell'Universitá di Cagliari*, vol. (III/II).

Giron J., Ginebra J., Riba A. (2004). Bayesian analysis of a multinomial sequence and homogeneity of literary style. *Doc. de Recerca 2004-01*. Dept. Estadistica i Investigacio Operativa, Universitat Politecnica de Catalunya. [Submitted for publication]

Givanel J. (1916). Les edicions gòtiques del Tirant lo Blanch en la Biblioteca de Catalunya. *Butlletí de a Biblioteca de Catalunya*, vol. (3): 58-72.

Goertz W. (1967). Zu Frage der Einheit des Tirant lo Blanc. *Romanistisches Jahrbuch*, vol. (18): 249-267.

Good I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, vol. (40): 237-264.

Good I.J. and Toulmin G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, vol. (43): 45-63.

Guia J. (1995). Introducció a la fraseologia del Tirant. *Afers*, vol. (20): 129-142.

Guia J. (1996). *De Martorell a Corella. Descobrint l'autor del Tirant lo Blanc*.

Harris B. (1982). Entropy. *Enclyclopedia of Statistical Sciences*, vol. (2): 512-516. Wiley.

Hauf A. (1993). Tirant lo Blanc: algunes qüestions que planteja la connexió corelliana. In Alemany R., Ferrando A. and Meseguer L. (Eds), *Actes del IX Col·loqui Internacional de Llengua i Literatura Catalanes (Alacant- Elx)*, vol. (2): 69-116

Hinkley D.V. (1970). Inference about a change-point in a sequence of random variables. *Biometrika*, vol. (57): 1-16.

Hinkley D.V. (1971). Inference in two phase regression. *J. Amer. Statist. Assoc.*, vol. (66): 736-743.

Holmes D.I. (1985). The Analysis of Literary Style - A Review. *Journal of the Royal Statistical Society, Series A*, vol. (148): 328-341.

Hubert P., Labbé C. and Labbé D. (2002). Segmentation automatique des corpus. Voyages de l'autre côté de J.-M. Le Clézio. In *Actes des JADT 2002 on line at* http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/tocJADT2002.htm

Lebart L., Salem A. and Berry L. (1998). *Exploring Textual Data*. Kluwer.

Manly B.F.J. (1994). Ecological statistics. In Patil G.P. and Rao C.R. (Eds), *Handbook of Statistics*, vol. (12): 307-376.

Margalef D.R. (1958). Information theory in ecology. *General Systems*, vol. (3): 36-71.

Martínez Martínez F. (1916). *Martín Joan de Galba, coautor del Tirant lo Blanch*. València.

Marinesco C. (1979). Nouvelles recherches sur Tirant lo Blanch. In *Miscel.lània Aramon i Serra*, vol. (I): 401-424.

Menéndez Pelayo M. (1943). *Orígenes de la novela*. CSIC: 392-403.

Moll F. (1933). Els refranys del Tirant lo Blanc.*Butlletí del diccionari de la llengua catalana*, vol. (15): 169-172.

Mosteller F. and Wallace D.L. (1984). *Applied Bayesian and Classical Inference; The Case of The Federalist Papers*, 2nd Edition. Springer-Verlag.

Nicolau d'Olwer L. (1961). Tirant lo Blanch: examen de algunas cuestiones. *Nueva revista de filología hispánica*, vol. (15): 131-154.

Patil G.P. and Taillie C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association*, vol. (77): 548-567.

Pettit A.N. (1989). Change-point problem. In *Encyclopedia of Statistical Sciences*, vol. (S): 26-31.

Pielou E.C. (1982). Diversity indices. In *Encyclopedia of Statistical Sciences*, vol. (2): 408-412.

Rao C.R. (1982). Analysis of diversity; A unified approach. *Statistical Decision Theory and Related Topics, III*, vol. (2): 233-250.

Riba A. (2002). *Homogeneitat d'estil en el Tirant lo Blanc (in catalan)*. [Unpublished phD thesis], Universitat Politècnica de Catalunya, Barcelona.

Riba A. and Ginebra J. (2003). Change-point estimation in a multinomial sequence and homogeneity of literary style. *Doc. de Recerca 2003-17*. Dept. Estadistica i Investigacio Operativa, Universitat Politecnica de Catalunya. [Submitted for publication].

Riquer M. (Ed.) (1947). *J. Martorell i M.J. Galba, Tirant lo Blanc*.

Riquer M. (Ed.) (1983). *Tirant lo Blanc*. Millors Obres de la Literatura Catalana, vols. (99-100), Edicions 62.

Riquer M. (1990). *Aproximació al Tirant lo Blanc*. Quaderns Crema.

Rubiera M.J. (1990). Tirant lo Blanc i els moros. *L'Aiguadolç*: 33-40.

Rubiera M.J. (1992). La mort de tirant i el triomf d'Hipòlit o la crisi del món cavalleresc vista per un cavaller en crisi. In *Acta del Convegno Internazionale: La cultura catalana tra l'umanesimo e il Barrocco*, Venèzia, Università degli studi di Venezia.

Sichel H.S. (1986a). Word frequency distributions and type-token characteristics. *Mathematical Scientist*, vol. (11): 45-72.

Sichel H.S. (1986b). Parameter estimation for a word frequency distribution based on occupancy theory. *Communications in Statistics, series A, theory and methods*, vol. (15): 935-949.

Simpson E.H. (1949). Measurement of diversity. *Nature*, vol. (163): 688.

Smith A.F.M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, vol. (63): 407-416.

Smith A.F.M. (1980). Change-point problems: approaches and applications. In Bernardo J.M., de Groot M.H, Lindley D.V. and Smith A.F.M. (Eds.), *Bayesian Statistics*: 83-98.

Thistedt R. and Efron B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika*, vol. (74): 445-455.

Vargas Llosa A.(1991). *Carta de batalla por Tirant lo Blanc*. Ed. Seix Barral.

Vargas Llosa A. (1993). Tirant lo Blanc: Las palabras como hechos. In *Actes del Symposion Tirant lo Blanc*: 587-604.

Vaeth A. (1918). *Tirant lo Blanch: a study of its authorship, principal sources and historical setting*: 93-94.

Wittlin C.J. (1990). Pistes per descobrir canvis introduïts per Martí Joan de Galba en el Tirant lo Blanch de Joanot Martorell. *Afers. Fulls de recerca i pensament*, vol. (5): 326-548.

Wittlin C.J. (1993). Dels manuscrits a l'edició: el Tirant, elaborat per Martorell el usant materials preexistents, revisat després en "valenciana prosa" per Galba. In *Actes del Sismposion Tirant lo Blanc*: 605-633.

Wolfe D.A. and Schechtman E. (1984). Nonparametric statistical procedures for the change point problem. *Journal of Statistical Planning and Inference*, vol. (9): 389-396.

EXIT : Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés

Mathieu Roche, Thomas Heitz, Oriane Matte-Tailliez, Yves Kodratoff

LRI – Université Paris-Sud – 91405 Orsay Cedex – France
{roche, oriane, yk}@lri.fr, thomashe@firtech.lri.fr

Abstract

The work presented in this paper is relative to the discovery of a significant terminology in specialized texts. Our approach, partly based on statistical methods extracts the terms in an iterative way. At first, the only terms looked for are binary. The binary terms detected during this first phase are included in the corpus, and the process is iteratively repeated in order to detect very long terms, that happen often to be the most significant terms, as our experience in molecular biology has clearly shown.

Résumé

Les travaux présentés dans cet article se rapportent à l'extraction de la terminologie à partir de textes spécialisés. Notre approche, en partie fondée sur des méthodes statistiques, consiste à extraire des termes de façon itérative. À la première itération, seuls des termes binaires sont recherchés. Ces termes sont alors réintroduits dans le corpus, et ceci est répété à chaque itération. De cette façon, nous pouvons extraire des termes de longueur croissante, particulièrement pertinents pour un domaine particulier, par exemple la biologie moléculaire.

Keywords: terminologie

1. Introduction

L'extraction de la terminologie du domaine à partir de textes est une tâche essentielle dans le but de construire des ontologies spécialisées. Cet article traite d'une étape importante préalable à celle de la terminologie : l'extraction des collocations (Halliday, 1976), c.-à-d. les groupes de mots. Les collocations jugées comme pertinentes sont celles qui peuvent être considérées comme des instances de concepts c'est-à-dire représenter des traces de concepts. Dans la suite, nous appellerons « termes » de telles collocations pertinentes car elles sont considérées par les experts du domaine comme des traces linguistiques de concepts de leur domaine.

Le but de nos travaux consiste à extraire les termes pertinents à partir de corpus spécialisés. Cependant, nous ne traiterons pas, dans cet article, les méthodes de regroupement des termes dans des concepts. Nous expliquerons le processus global d'extraction de la terminologie du domaine que nous proposons : le système EXIT (EXtraction Itérative de la Terminologie). Une des caractéristiques essentielles de ce système provient de la méthode itérative et statistique qui le caractérise, comme nous allons le montrer dans la suite.

Nous travaillons actuellement à partir de quatre corpus de tailles, de langues et de spécialités différentes. Ainsi, nous étudions un corpus en anglais composé de 6119 résumés d'articles relatifs à la biologie moléculaire (9424 Ko). Le deuxième corpus étudié est composé de 100 textes (369 Ko) d'introductions d'articles traitant de la fouille de données. Le troisième corpus

est issu du domaine des ressources humaines (Compagnie PerformanSe). Ce dernier a une taille de 3784 Ko et a été rédigé par un psychologue qui sert d'expert pour ce corpus. Enfin, nous disposons également d'un corpus de Curriculum Vitæ, d'une taille de 2470 Ko qui contient 1144 CVs (Groupe VediorBis). Dans les expérimentations que nous décrirons dans cet article, nous traiterons seulement le corpus des ressources humaines et le corpus de biologie moléculaire[1] qui sont les plus significatifs en terme de taille. Avant d'expliquer notre méthode d'extraction de la terminologie, nous résumons l'état de l'art du domaine dans la section suivante.

2. État de l'art

De multiples approches de recherche terminologique ont été développées afin d'extraire les termes pertinents à partir d'un corpus. Nous ne traiterons pas ici les approches d'aide à la structuration et au regoupement conceptuel des termes qui sont détaillés dans (Aussenac-Gilles et Bourigault, 2003).

Les méthodes d'extraction de la terminologie sont fondées sur des méthodes statistiques ou syntaxiques. TERMINO (David et Plante, 1990) est un outil précurseur qui s'appuie sur une analyse syntaxique afin d'extraire les termes nominaux. Cet outil effectue une analyse morphologique à base de règles, suivie de l'analyse des collocations nominales à l'aide d'une grammaire. Les travaux de (Smadja, 1993) (XTRACT) s'appuient sur une méthode statistique. XTRACT extrait, dans un premier temps, les collocations binaires situées dans une fenêtre de dix mots. Les collocations binaires sélectionnées sont celles qui dépassent d'une manière statistiquement significative la fréquence due au hasard. L'étape suivante consiste à extraire les collocations plus générales (collocations de plus de deux mots) contenant les collocations binaires trouvées à la précédente étape. ACABI (Daille, 1994) effectue une analyse linguistique afin de transformer les collocations nominales en termes binaires. Ces derniers sont ensuite triés selon des mesures statistiques. Afin d'extraire la terminologie du domaine, LEXTER (Bourigault et Jacquemin, 1999 ; Aussenac-Gilles et Bourigault, 2000) s'appuie uniquement sur une analyse syntaxique, contrairement au système ACABI, qui est fondé sur une méthode statistique. La méthode consiste à extraire les syntagmes nominaux maximaux. Ces syntagmes sont alors décomposés en termes de « têtes » et d'« expansions » à l'aide de règles grammaticales. Les termes sont alors proposés sous forme de réseau organisé en fonction de critères syntaxiques.

La section suivante présente l'approche globale que nous proposons et qui s'appuie sur des méthodes statistiques avec une approche itérative.

3. EXIT : Une approche itérative pour extraire la terminologie du domaine

L'approche globale que nous proposons avec EXIT se déroule en trois phases.

Comme dans de nombreux travaux, nous nous sommes intéressés, dans la première phase, à l'extraction de la terminologie nominale. L'originalité de notre méthode vient du procédé itératif que nous proposons. Cette méthode consiste, à chaque itération, à former des termes binaires (ou ternaires pour les termes prépositionnels). Chaque terme est sélectionné selon une mesure statistique (voir section 4.2.). Par exemple, sur le corpus des ressources humaines, à la première itération, nous avons extrait le terme binaire « travail administratif ». Ces termes sont réintroduits dans le corpus avec des traits d'union afin qu'ils soient reconnus comme des mots à part entière. Nous pouvons ainsi effectuer une nouvelle recherche terminologique à partir de ce corpus avec prise en compte de la terminologie du domaine acquise à l'étape précédente.

[1] Quelques centaines de termes que nous avons extraits à partir du corpus de biologie moléculaire sont consultables à l'adresse : http ://www.lri.fr/ia/Genomics/

Par exemple, à l'aide du terme binaire « travail administratif » trouvé à la première itération, nous pouvons extraire le terme « responsabilité de travail-administratif » à la deuxième itération. Notre méthode permet alors de détecter des termes très spécifiques (composés de plusieurs mots). Ceci est essentiel, par exemple en biologie moléculaire, où les termes les plus pertinents sont les termes composés de nombreux mots. Nous démontrerons dans la section 4.5., la terminaison de notre algorithme. De plus, comme nous le montrerons dans la section 4.6., nous avons ajouté un certain nombre de paramètres afin d'augmenter la pertinence des termes extraits.

La deuxième phase de notre travail consiste à extraire la terminologie verbale. Nous nous sommes attachés à extraire les termes du type « verbe-objet » et « sujet-verbe » comme dans les travaux de Smadja (1993). Par exemple, grâce à la terminologie nominale extraite dans la première phase, nous avons alors pu obtenir le terme verbal « prendre responsabilité-de-travail-administratif ». De même que pour les termes nominaux, les termes verbaux sont ordonnancés selon une mesure statistique.

Enfin, dans la dernière phase, nous avons utilisé FASTR (Jacquemin, 1996) afin de déterminer les termes variants à partir des termes nominaux extraits mais également à partir des termes verbaux.

La figure 1 illustre le schéma global de l'ensemble de notre processus. Le détail de chacune de ces phases est donné dans les sections suivantes.

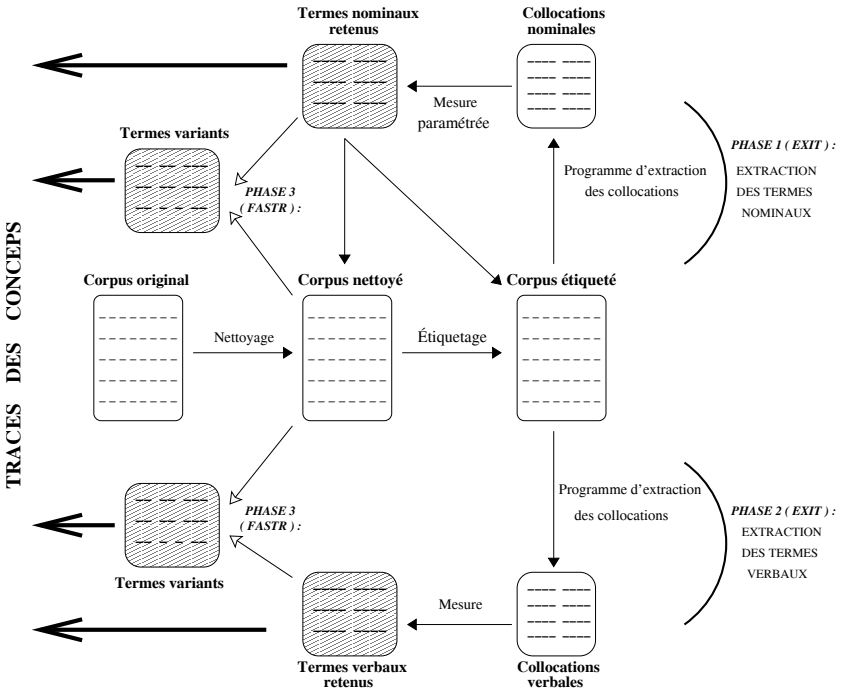


Figure 1. Schéma global de la recherche terminologique dans les textes avec EXIT.

4. Extraction de la terminologie nominale

4.1. *Nettoyage et étiquetage*

La première étape de notre travail a consisté à normaliser le corpus. Pour effectuer ce travail, nous supprimons les éléments susceptibles d'engendrer des erreurs avec les outils automatiques d'analyse des textes (étiqueteurs, analyseurs syntaxiques, etc.). Ainsi nous avons supprimé le nom des auteurs des articles que nous traitons, le nom des laboratoires, etc. Avec l'expert du

domaine, nous avons également établi des règles afin d'uniformiser le vocabulaire employé. Par exemple, sur le corpus de biologie moléculaire, nous avons remplacé les termes « carboxyl terminal », « carboxyl termini », « COOH-terminal », « C-terminal », etc. par « C-term ». L'expert du domaine a également construit une règle lexicale afin de repérer les formules dans le corpus de biologie moléculaire (par exemple, « GAL4 », « RAP1 », « FK506 », « Mcm1 », etc.). Ces formules sont souvent associées à un nom significatif pour le domaine (« gene », « protein », « promoter », « cell », etc.). Une règle de nettoyage simple (ou de pré-terminologie) a alors consisté à rassembler les formules et les mots significatifs par un trait d'union (« formule-nom_significatif »). Une telle tâche permet à l'étiqueteur de Brill que nous allons utiliser à l'étape suivante, de reconnaître de tels regroupements comme des mots à part entière. D'autres règles du même type ont été construites à l'aide de la connaissance du domaine fournie par l'expert

La deuxième étape de notre travail utilise l'étiqueteur de Brill (1994) afin d'apposer une étiquette grammaticale à chacun des mots du corpus. Nous avons également ajouté quelques règles lexicales et contextuelles afin d'enrichir le lexique standard de Brill qui possédait bon nombre de mots inconnus. L'enrichissement du lexique est essentiel, en particulier sur un domaine très spécialisé tel que la biologie moléculaire pour lequel plus 70% des mots ne sont pas présents dans le lexique standard de l'étiqueteur de Brill. Avec l'ajout de règles lexicales et contextuelles établies par l'expert, tous les mots du corpus de biologie moléculaire sont alors reconnus par l'étiqueteur.

L'étape suivante a consisté à extraire les collocations binaires (ou ternaires pour les collocations prépositionnelles), c'est-à-dire les mots voisins ayant une étiquette spécifique. Nous nous sommes ainsi intéressés aux collocations du type *nom-nom*, *adjectif-nom*, *nom-adjectif*[2], *nom-préposition-nom* et *formule-nom*[3] car elles sont les plus porteuses de traces de concepts.

La section suivante présente la manière dont nous allons ordonnancer ces collocations.

4.2. *Mesure utilisée avec* EXIT

Après avoir extrait l'ensemble des collocations binaires ou ternaires, nous ordonnançons ces dernières selon une mesure statistique. Plusieurs mesures sont décrites dans la littérature. Nous avons testé les versions traitées dans Jacquemin (1997), Daille *et al.* (1998) et Lenca *et al.* (2003). La comparaison des résultats obtenus avec 11 mesures est détaillée dans Roche *et al.* (2003). Dans cet article, nous présentons deux mesures typiques pour l'extraction de la terminologie : l'information mutuelle et le rapport de vraisemblance.

La première mesure que nous décrivons correspond à l'information mutuelle (Church et Hanks, 1990). Cette mesure est définie par la formule (1) :

$$IM(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

L'information mutuelle calcule une certaine forme d'indépendance des mots x et y composant la collocation. Une telle mesure a tendance à extraire des termes rares et peu fréquents.

La deuxième mesure que nous citerons dans cet article correspond au rapport de vraisemblance (Dunning, 1993). Comme le précisent Daille *et al.* (1998) ainsi que Xu *et al.* (2002), cette deuxième mesure est particulièrement bien adaptée pour l'extraction de la terminologie. Nous

[2] Pour le corpus français uniquement.
[3] L'étiquette « formule » est une étiquette spécifique au corpus de biologie moléculaire.

définissons le rapport de vraisemblance de la manière suivante à l'aide d'une table de contingences associée à chaque couple de mots (x_i, x_j) comme ci-dessous :

	x_j	$x_{j'}$ avec $j' \neq j$
x_i	a	b
$x_{i'}$ avec $i' \neq i$	c	d

Les valeurs a, b, c et d définissent les occurences des couples et $a + b + c + d = N$ est le nombre total d'occurences des couples trouvés.

On définit alors le rapport de vraisemblance par la formule (2) :

$$\begin{aligned} RV(x_i, x_j) &= a\log(a) + b\log(b) + c\log(c) + d\log(d) - (a+b)\log(a+b) \\ &\quad -(a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) \\ &\quad +N\log(N) \end{aligned} \tag{2}$$

Ces deux mesures affectent un score aux collocations (x, y). Le rapport de vraisemblance prend en compte les collocations formées avec un seul des mots x et y ainsi que les collocations formées avec aucun de ces mots x et y. A contrario, l'information mutuelle s'appuie uniquement sur l'indépendance de chacun des mots x et y composant la collocation.

4.3. *Expérimentations*

Afin de comparer les différentes mesures de la littérature, nous nous appuyons sur la mesure de Précision et son complément la courbe d'élévation (« lift chart »), qui sont les seules mesures d'évaluation dont nous puissions disposer dans un cadre d'apprentissage non-supervisé. La précision, qui donne la proportion de collocations correctes parmi les collocations extraites, est définie par la formule (3) :

$$Précision = \frac{\text{nombre de collocations extraites pertinentes}}{\text{nombre de collocations extraites}} \tag{3}$$

Comme le précisent Aussenac-Gilles et Bourigault (2000), selon l'utilisation que l'on compte faire de la terminologie (indexation, ontologie, etc.), l'évaluation de la pertinence des termes peut différer. Pour notre part, nous jugeons qu'une collocation est pertinente si elle représente une trace de concept.

La courbe d'élévation donne la précision en fonction de la proportion de collocations proposées à l'expert. Précisons qu'il est impossible de connaître le nombre total de collocations pertinentes des corpus, c'est la raison pour laquelle nous ne calculerons ni le Rappel ni la courbe ROC (Ferri *et al.*, 2002).

À titre d'exemple, la figure 2 présente la courbe d'élévation des deux mesures décrites dans la section précédente, appliquées au corpus de biologie moléculaire. Ainsi, la figure 2 montre que la précision est toujours plus importante avec le rapport de vraisemblance suivant le nombre de collocations proposées à l'expert. Notons que les autres corpus donnent des résultats similaires.

Plus généralement, rappelons que parmi la dizaine de mesures testées dans Roche *et al.* (2003), le rapport de vraisemblance est la mesure qui a le meilleur comportement pour l'extraction de la terminologie.

Nous précisons que le critère consistant à simplement classer les collocations selon leur nombre d'occurences peut également se révéler efficace. Cependant, il est nécessaire d'effectuer un classement des collocations ayant le même nombre d'occurences en utilisant une mesure statistique.

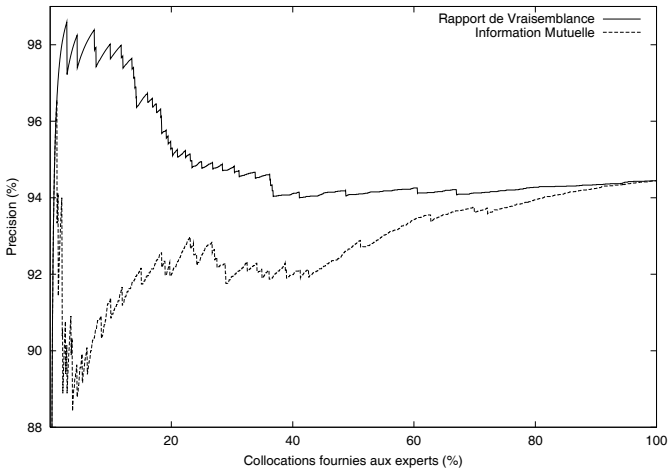


Figure 2. Courbes d'élévation propres au corpus de biologie moléculaire avec la relation adjectif-nom et un élagage à 4 avec l'information mutuelle et le rapport de vraisemblance

4.4. *Sélection des collocations*

Une fois les mesures statistiques effectuées, l'ensemble ordonné de ces collocations (ou un sous-ensemble d'entre elles selon le choix de l'expert) est proposé à l'expert. L'expert les classe alors en quatre catégories.

1. La première catégorie correspond aux collocations reconnues comme pertinentes par l'expert. Nous noterons $\mathcal{L}_1^i$ cet ensemble de collocations de l'itération i.

2. La deuxième catégorie correspond aux collocations reconnues comme non pertinentes par l'expert. Nous noterons $\mathcal{L}_2^i$ ces collocations de l'itération i.

3. La troisième catégorie correspond aux collocations qui ne sont pas pertinentes mais qui pourraient être pertinentes lors de la formation d'un terme plus long et donc plus spécifique. Nous noterons $\mathcal{L}_3^i$ ces collocations de l'itération i.

4. La dernière catégorie est composée des collocations non évaluées par l'expert. Nous noterons $\mathcal{L}_4^i$ cet ensemble de collocations relatives à l'itération i.

Pour faciliter, le travail de l'expert, nous avons développé une interface graphique en Java (voir figure 3). Cette interface permet de sélectionner les collocations selon les catégories décrites. Nous précisons que pour faciliter le travail de l'expert, nous proposons à l'utilisateur de visualiser les contextes (phrases) dans lesquels apparaissent les collocations.

De plus, nous proposons à l'expert la possibilité de valider certaines collocations si elles respectent des règles lexicales que l'expert peut construire à l'aide de l'interface graphique. Ainsi, une partie des collocations est validée de façon semi-automatique comme collocations pertinentes ou non pertinentes.

Enfin, le système que nous avons développé utilise le principe de l'élagage progressif suivant les itérations. L'élagage correspond au fait de proposer à l'expert uniquement les collocations qui apparaissent un nombre de fois minimum dans le corpus. Afin de sélectionner les termes de plus en plus significatifs à chaque itération, l'expert peut effectuer un élagage plus important selon les itérations. Nous appelons ce principe un élagage progressif.

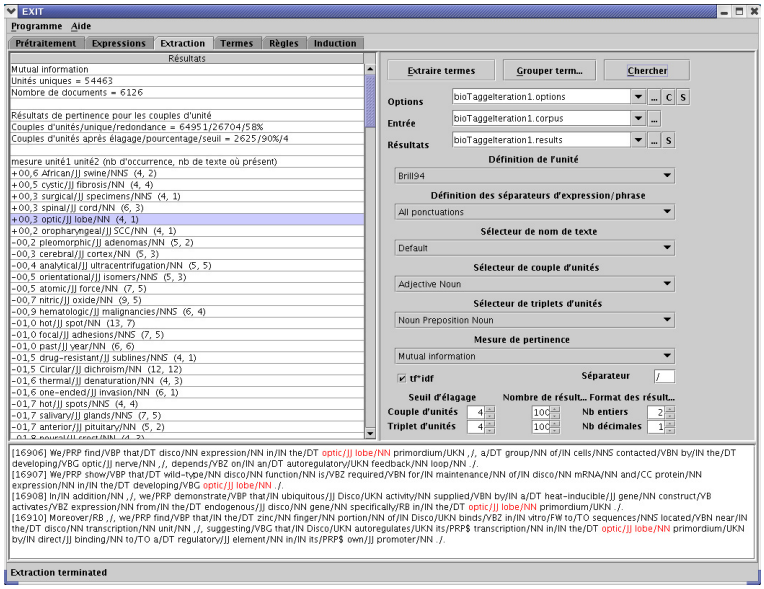


Figure 3. Capture d'écran du logiciel EXIT. *La partie gauche présente les différentes collocations classées selon une mesure statistique. La partie droite présente les différents paramètres (types de collocations à extraire, mesures à utiliser, élagage, etc.). La partie inférieure présente les contextes (phrases) dans lesquels nous trouvons les collocations sélectionnées par l'expert*

4.5. *Terminaison de l'algorithme*

Pour démontrer la terminaison de notre algorithme itératif de recherche terminologique, nous en présentons ci-dessous quelques caractéristiques.

- Soit N_i le nombre total de collocations extraites à l'itération i.
- Soit n_i le nombre total de collocations extraites après un élagage λ_i à la $i$ème itération.
- Soit α_i le nombre de collocations retenues dans les listes $\mathcal{L}_k^i$ où $1 \leq k \leq 3$ (voir section 4.4) parmi les n_i collocations extraites à l'itération i.
  Ainsi, $\alpha_i = card(\bigcup_{k=1}^{3} \mathcal{L}_k^i)$. On suppose qu'à chaque itération, au moins une collocation est retenue, c.-à-d. $\forall i, \alpha_i > 0$.

Bien entendu, $\forall i, \alpha_i \leq n_i \leq N_i$. S'il n'y a pas d'élagage ($\lambda_i = 1$) alors $n_i = N_i$.

L'algorithme s'arrête lorsque nous obtenons une liste vide de collocations après élagage. Ainsi, le nombre d'itérations de notre algorithme revient à déterminer j tel que $n_{j+1} = 0$.

Dans le meilleur des cas, cas trivial pour démontrer la terminaison de notre algorithme, nous sélectionnons toutes les collocations de la première itération parmi les listes $\mathcal{L}_k^1$ où $1 \leq k \leq 3$. Dans ce cas $j = 1$ et on a $n_1 = \alpha_1$.

Dans le pire des cas, on suppose qu'il n'y a pas d'élagage progressif (c.-à-d. $\forall (i, j), \lambda_i = \lambda_j$) et que l'on ne sélectionne qu'une seule collocation par itération ($\forall i, \alpha_i = 1$). Dans ce cas, l'algorithme se termine après $j = n_1$ itérations. Ce nombre n_1 étant nécessairement fini, la terminaison de notre algorithme est donc assurée dans le pire des cas.

Notons qu'ici, nous supposons qu'à l'itération i, lorsqu'un seul terme est validé parmi les n_i collocations extraites, à l'itération suivante, nous avons $n_i - 1$ collocations qui sont extraites. Ainsi, nous n'avons pas pris en compte le cas rare où la sélection d'un terme permet de construire une nouvelle collocation binaire qui ne pouvait être établie lors des itérations précédentes (par exemple, avec la structure « nom adjectif1 adjectif2 » où seule la collocation « nom adjectif1 » est proposée à l'itération i puis « nom-adjectif1 adjectif2 » à l'itération $i + 1$). La garantie de

terminaison de l'algorithme reste cependant vérifiée car le nombre de termes formés dans ce cas est nécessairement borné (la borne correspond alors au nombre fini de mots moins 1 présents dans les phrases rencontrant un tel cas).

Dans le cas général, et s'il n'y a pas d'élagage progressif suivant les itérations, l'algorithme se termine après j itérations, où j vérifie $\sum_{i=1}^{j} \alpha_i = n_1$. S'il y a un élagage progressif, l'algorithme se termine avec j_0 tel que $j_0 \leq j$.

4.6. *Paramètres supplémentaires*

Dans cet article, nous détaillons deux paramètres qui sont utilisés pour l'extraction de la terminologie. Nous présentons ici les paramètres que nous avons modifiés comparativement aux versions présentées dans Roche (2003).

Un des paramètres que nous avons développé consiste à sélectionner automatiquement les termes déjà validés par les auteurs du corpus. Ce principe consiste à considérer, lors de la première itération, les termes du corpus qui possèdent un trait d'union (par exemple, « data-mining ») comme des termes pertinents. Les mêmes collocations que l'on trouve sans trait d'union (exemple, « data mining ») sont également reconnues comme des termes valides et sont introduits automatiquement dans la liste $\mathcal{L}_1^1$. Ce principe revient à utiliser l'expertise implicite de certains auteurs. Bien entendu, l'utilisateur peut éliminer de $\mathcal{L}_1^1$ tout terme qu'il juge non pertinent. Dans la phase de nettoyage (voir section 4.1), nous avons, par exemple, identifié des termes de type « formule nom_spécifique » que nous avons remplacés par le regroupement « formule-nom_spécifique ». Ces termes sont *a priori* pertinents car ils ont été détectés semi-automatiquement grâce à des règles lexicales construites par l'expert du domaine durant la phase de nettoyage. Le paramètre que nous décrivons ici permet alors de placer automatiquement de tels termes dans la liste $\mathcal{L}_1^1$ des termes valides.

Un autre paramètre essentiel que nous avons également utilisé consiste à priviligier les collocations qui apparaissent dans des textes différents, jugées comme davantage représentatives du domaine, comparativement aux collocations qui sont présentes dans peu de textes.

Pour cela, nous allons nous appuyer sur la pondération classique dans le domaine de la recherche d'informations qui est le tf $\times$ idf (term frequency - inverse document frequency). Cette pondération est donnée par la formule (4).

$$w_{ij} \quad = \quad \text{tf}_{ij} \times \log_2 \frac{N}{n} \tag{4}$$

– w_{ij} est le poids du terme T_j dans le document D_i ;
– tf_{ij} est la fréquence du terme T_j dans le document D_i ;
– N est le nombre de documents dans la collection ;
– n est le nombre de documents où le terme T_j apparaît au moins une fois.

Cette formule combine l'importance des termes pour un document (tf) et le pouvoir de discrimination de ce terme (idf). Ainsi, un terme qui a une valeur de tf $\times$ idf élevée doit être à la fois important dans ce document et il doit apparaître peu dans les autres documents. C'est le cas où un terme correspond à une caractéristique importante et unique d'un document. Une telle mesure est souvent utilisée en recherche documentaire. Dans notre cas, nous souhaitons obtenir une réponse inverse. Ainsi, nous souhaitons privilégier les collocations fréquentes qui apparaissent dans de nombreux documents. Pour cela, nous avons modifié la formule classique du tf $\times$ idf de la manière suivante (5) :

$$w'_{ij} \quad = \quad \frac{\text{tf}_{ij}}{\log_2 \frac{N}{n}} \tag{5}$$

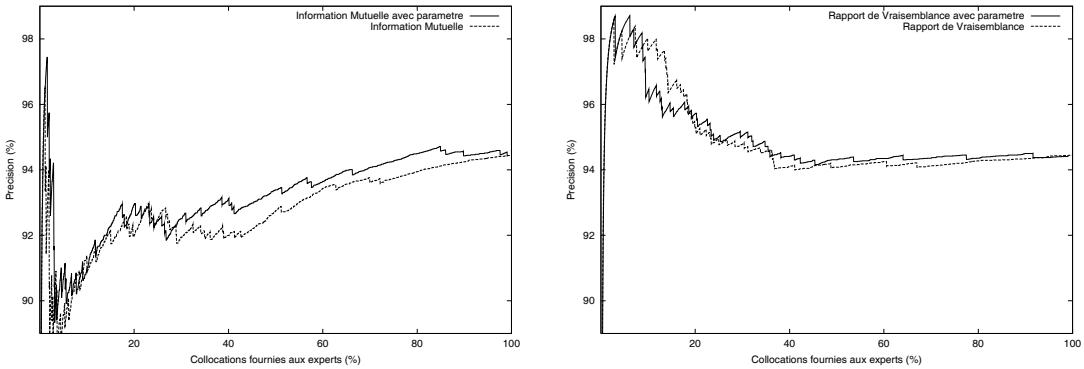


Figure 4. Information mutuelle (a) et Rapport de Vraisemblance (b) avec paramètre

Nous avons estimé l'amélioration de la précision obtenue en remplaçant le nombre d'occurences des collocations et des mots par la pondération décrite précédemment. Globalement, la précision est améliorée en utilisant notre nouveau paramètre. À titre d'exemple, avec l'information mutuelle appliquée à la relation adjectif-nom du corpus de biologie moléculaire, la précision est sensiblement améliorée (voir figure 4.a). Pour le rapport de vraisemblance (voir figure 4.b), la précision est globalement meilleure en utilisant notre paramètre. Cependant, si nous fournissons aux experts un faible pourcentage de collocations (moins de 20%), l'utilisation de notre paramètre n'améliore pas toujours les résultats en terme de précision.

5. Extraction de la terminologie verbale

Dans cette étape, nous nous sommes intéressés à l'extraction des collocations verbales de type « verbe-objet » et « sujet-verbe » comme dans les travaux de Smadja (1993). Cette extraction des collocations s'effectue grâce à des règles contextuelles consistant à trouver les collocations binaires formées avec les étiquettes relatives aux noms et aux verbes. La différence entre collocations verbales et collocations nominales tient au fait que certaines collocations verbales peuvent se révéler erronées. Ainsi, nous avons estimé que le taux des collocations verbales syntaxiquement correctes est à plus de 80%. Ces collocations binaires sont ensuite ordonnancées selon le score donné par le rapport de vraisemblance.

6. Extraction des termes variants

Nous avons évalué les termes variants donnés par FASTR (Jacquemin, 1996). Ainsi, Jacquemin (1996) définit, par exemple, trois types de termes variants :

– $X_2X_3 \mapsto X_2X_4C_5X_3$. Par exemple, avec le corpus de biologie moléculaire, nous avons *Ty1 transposition* $\mapsto$ *Ty1 transcription and transposition*
– $X_2X_3 \mapsto X_2X_4X_3$. Par exemple, *close proximity* $\mapsto$ *close chromosomal proximity*
– $X_2X_3 \mapsto X_3P_4X_5X_2$. Par exemple, *protein sequence* $\mapsto$ *sequence for dimeric protein*

Nous avons effectué une évaluation des termes variants du corpus de biologie moléculaire. Sur 1200 termes variants donnés par FASTR à partir des termes nominaux extraits à la première itération, 85% des termes ont été jugés pertinents. Cette expertise donne des résultats similaires à ceux cités dans Jacquemin (1996) correspondant à un corpus médical. Avec ce corpus, les termes variants obtenus atteignaient également une précision de près de 85%.

7. Conclusions et perspectives

L'approche que nous proposons dans cet article est une approche globale pour extraire la terminologie à partir de corpus spécialisés. Notre méthode s'appuie essentiellement sur une méthode itérative et des mesures statitistiques. Nous avons validé notre approche sur des corpus de langues, de tailles et de domaines différents.

La méthodologie décrite ici a deux perspectives immédiates. Premièrement, l'expert examine de nombreuses phrases contenant les termes détectés afin de les valider. Lorsqu'il juge que tous les termes intéressants pour lui ont été trouvés, il peut alors conserver ces phrases, créant ainsi un sous-corpus « parfaitement » étiqueté sur lequel des mesures ROC et de Rappel peuvent être effectuées. Cette option est implantée dans notre système mais n'a pas encore été systématiquement utilisée par les experts.

Deuxièmement, la base étiquetée constitue un ensemble d'apprentissage (bruité en général mais non bruité pour un sous-corpus) à partir duquel un outil d'induction automatique peut créer de nouvelles règles dont la pertinence pourra être jugée par l'expert. Le problème majeur posé par cette approche est celui de la définition des descripteurs de ce qui est un « bon » terme. Nos résultats préliminaires montrent que nous aurons besoin d'outils semblables à ceux utilisés pour la détection des traces linguistiques de concepts, à savoir « l'induction extensionnelle » définie dans Kodratoff (2004).

Références

Aussenac-Gilles N. et Bourigault D. (2000). The Th(IC)2 Initiative : Corpus-Based Thesaurus Construction for Indexing WWW Documents. In *Proceedings of the EKAW'2000 Workshop on Ontologies and Texts*, vol. (51).

Aussenac-Gilles N. et Bourigault D. (2003). Construction d'ontologies à partir de textes. In *Actes de TALN03*, vol. (2) : 27-47.

Bourigault D. et Jacquemin C. (1999). Term Extraction + Term Clustering : An Integrated Platform for Computer-Aided Terminology. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL '99)* : 15-22.

Brill E. (1994). Some advances in transformation-based part of speech tagging. In *AAAI*, vol. (1) : 722-727.

Church K. W. et Hanks P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, vol. (16) : 22-29.

Daille B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de Doctorat en Informatique Fondamentale. Université Paris 7.

Daille B., Gaussier E. et Langé J. (1998). An Evaluation of Statistical Scores for Word Association. In Ginzburg J., Khasidashvili Z., Vogel C., Levy J.-J. et Vallduvi E. (Eds), *The Tbilisi Symposium on Logic, Language and Computation : Selected Papers*. CSLI Publications : 177-188.

David S. et Plante P. (1990). De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Québec*, vol. (3) : 140-154.

Dunning T.E. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol. (19/1) : 61-74.

Ferri C., Flach P. et Hernández-Orallo J. (2002). Learning decision trees using the area under the ROC curve. In *Proceedings of the 9th International Conference on Machine Learning, ICML'02* : 139-146.

Halliday M.A.K. (1976). *System and Function in Language*. Oxford University Press.

Jacquemin C. (1996). A symbolic and surgical acquisition of terms through variation. In Wermter S., -Riloff E. et Scheler G. (Eds), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing* : 425-438.

Jacquemin C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale. Université de Nantes.

Kodratoff Y. (2004). Induction Extentionnelle : définition et application à l'acquisition de concepts à partir de textes. *RNTI*, (numéro spécial EGC04) [à paraître].

Lenca P., Meyer P., Picouet P., Vaillant B. et Lallich S. (2003). Critères d'évaluation des mesures de qualité des règles d'association. *RNTI*, (n° spécial "entreposage et fouille des données"). *CEPADUES* : 123-134.

Roche M. (2003). Extraction paramétrée de la terminologie du domaine. *RSTI série RIA-ECA*, (numéro spécial EGC03), vol. (17) : 295-306.

Roche M., Matte-Tailliez O., Azé J. et Kodratoff Y. (2003). Extraction de la Terminologie du Domaine : Étude de Mesures sur un Corpus Spécialisé Issu du Web. In *Actes des Journées Francophones de la Toile 2003* : 279-288.

Smadja F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics*, vol. (19) : 143-177.

Xu F., Kurz D., Piskorski J. et Schmeier S. (2002). A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping. In *Proceedings of LREC 2002*.

Il processo di codifica dei dati testuali dell'indagine Multiscopo "Uso del tempo"

Maria Clelia Romano, Tania Cappadozzi[1]

ISTAT – SDS/2 – Via Ravà 150 – 00142, Roma – Italy
romano@istat.it

Abstract

During the performance of Time Use surveys the climax is the coding phase of daily activities performed by respondents and freely expressed in the diary. In this paper will be described the main choices adopted by Italian National Institute for Statistics during the planning of Time Use Survey 2002-2003, specially with reference to the textual data collected treatment. Two important changes have been introduced: first, the data entry of textual data reported in the diaries, that allowed to build a very innovating database for dimension and content; second, the implementation of an assisted coding system that allowed to monitor whole coding process and to better the quality of final data.

Riassunto

Una fase cruciale nella realizzazione delle indagini sull'uso del tempo è rappresentata dalla codifica delle attività giornaliere svolte dagli intervistati ed espresse in forma libera nei diari. Nel presente lavoro descriveremo le principali scelte adottate dalll'Istituto Nazionale di Statistica nel corso della riprogettazione dell'indagine 2002-2003, con riferimento al trattamento dei dati testuali raccolti. In particolare, due importanti novità sono state introdotte in questa fase: la registrazione dei dati testuali riportati nei diari, che ha condotto alla costruzione di un archivio dati per dimensioni e contenuto di grande rilevanza, e l'implementazione di un sistema di codifica assistita che ha consentito di monitorare l'intero processo di codifica e di migliorare la qualità del dato finale.

Keywords: time use, textual data, daily activities, hierarchical classification, coding list adaptations.

1. L'indagine Istat Multiscopo "Uso del tempo" 2002-03

Tra aprile 2002 e marzo 2003 si è svolta la rilevazione Istat Multiscopo sulle famiglie "Uso del tempo" 2002-03. Nel corso di tale indagine è stato intervistato un campione di 21.075 famiglie per un totale di circa 55.000 individui.

Gli strumenti di rilevazione utilizzati sono quattro: un questionario per intervista individuale, un questionario per intervista familiare, un diario settimanale destinato alle persone di 15 anni e più mirato a rilevare le ore di lavoro retribuito svolte nel corso di una data settimana, e un diario giornaliero destinato alle persone di 3 anni e più e destinato a rilevare specificamente l'uso del tempo quotidiano.

Lo strumento principe di questa indagine è ovviamente rappresentato dal diario giornaliero, che costituisce una fonte preziosa di informazioni ad elevato livello di dettaglio sull'organizzazione dei tempi individuali, ed apre, essendo l'unità di rilevazione la famiglia, la possi-

[1] Il lavoro è frutto della collaborazione delle due autrici. Tuttavia la stesura dei paragrafi 1, 2, 3, 7 è da attribuire a Maria Clelia Romano e quella dei paragrafi 4, 5, 6 a Tania Cappadozzi.

bilità di un'analisi interindividuale dell'uso del tempo, poiché consente di rilevare le varie combinazioni tra i tempi di tutti i componenti della famiglia.

Infatti, attraverso la compilazione del diario è possibile conoscere il modo in cui ciascun rispondente ripartisce 24 ore (divise in 144 intervalli di 10 minuti) tra le varie attività giornaliere, gli spostamenti, i luoghi frequentati e le persone con cui le ha trascorse (vedi fig. 1). Si tratta cioè di un'informazione che presenta un livello di dettaglio estremamente elevato, non comparabile con quelle ricavate dai tradizionali questionari a domande fisse.

L'indagine "Uso del tempo" rientra inoltre in un progetto Eurostat di armonizzazione delle indagini "Time Use" che nato nel 1994 è giunto alla formulazione di Linee Guida (Eurostat, 1999), che se rispettate nella progettazione delle indagini nazionali, consentono di perseguire l'obiettivo della comparazione internazionale dei risultati. Secondo tali linee guida e secondo quanto realizzato nella maggior parte dei paesi europei, l'uso del tempo rilevato attraverso lo schema del diario giornaliero viene poi classificato secondo un analitico elenco di codici numerici.

La descrizione delle attività e dei luoghi, dunque, pur essendo lasciata alla libera capacità di espressione di ciascun rispondente viene in seguito sottoposta ad una classificazione. In Italia, il sistema di classificazione adottato, pur partendo da quello suggerito da Eurostat, e conservando pertanto le conseguenti possibilità comparative, è stato integrato ed adattato alle specifiche esigenze nazionali.

	Che cosa sta facendo? Indichi le attività che svolge ad intervalli di tempo di 10 minuti!	Che cos'altro sta facendo? Indichi l'attività contemporanea più importante	Dove si trova o come si sta spostando?	E da solo o con persone che conosce? Non risponda se sta dormendo, lavorando o se sta a scuola				
				Da solo	Con familiari conviventi meno di 10 min	Con familiari conviventi più di 10 min	Con altri familiari non conviventi	Con altre persone che conosce
07.00 - 07.10	Mi sono svegliata		A casa mia			X		
07.10 - 07.20	Mi sono lavata			X				
07.20 - 07.30	Ho svegliato mio figlio				X			
07.30 - 07.40	Ho preparato la colazione	Ho ascoltato la radio						
07.40 - 07.50	Ho fatto colazione	Ho letto il quotidiano						
07.50 - 08.00	Ho lavato i piatti	Ho ascoltato la radio						
08.00 - 08.10	Ho rifatto i letti	" "						
08.10 - 08.20	Ho portato il figlio al nido	Ho parlato con figlio	A piedi		X			
08.20 - 08.30	Sono andata al lavoro		In autobus	X				
08.30 - 08.40	Lavoro (unica occupazione)		In ufficio					
08.40 - 08.50			" "					
08.50 - 09.00			" "					
09.00 - 09.10			" "					
09.10 - 09.20			" "					
09.20 - 09.30			" "					
09.30 - 09.40			" "					
09.40 - 09.50			" "					
09.50 - 10.00			" "					

Figura 1. Lo schema del diario giornaliero

Obiettivo di questo lavoro è descrivere le principali scelte effettuate dall'Istituto Nazionale di Statistica nella progettazione dell'indagine "Uso del tempo" 2002-03, con particolare riferimento alle strategie innovative adottate per la classificazione delle informazioni testuali rilevate tramite i diari giornalieri. In particolare, viene descritta l'esperienza della codifica assistita e dei principali vantaggi che da tale scelta sono derivati, in termini di possibilità di monitoraggio del processo di codifica e di miglioramento della qualità dei dati.

2. La codifica: caratteristiche e problematiche

Buona parte della riuscita di questa indagine dipende dall'accuratezza con cui viene svolto il lavoro di codifica delle informazioni contenute nei diari giornalieri. Solo l'esaustività del sistema di classificazione e la corretta applicazione dello stesso possono garantire elevati standard di risultato, oltre che possibilità comparative a livello internazionale.

Il sistema di classificazione Eurostat ha una struttura gerarchica, fino a un massimo di quattro livelli, aperto alla possibilità di inserire codici specifici di interesse nazionale. Dieci sono le categorie di primo livello, individuate per classificare le attività in base agli obiettivi e al contesto (per esempio: Lavoro, Studio, Cura della casa e gestione della famiglia, Vita sociale etc.). Di seguito si riporta un esempio della struttura gerarchica:

3. CURA DELLA PROPRIA CASA E DELLA PROPRIA FAMIGLIA

3.1 CUCINARE, LAVARE E RIORDINARE LE STOVIGLIE

3.1.1 Preparare i pasti, le merende, la colazione

3.1.2 Preparare torte, biscotti, pane e altri alimenti con farina

3.1.3 Lavare le stoviglie

3.1.3.1 Apparecchiare e sparecchiare la tavola

3.1.4 Preparare conserve, confetture, congelare, imbottigliare, ecc.

.

.

. *(Altre categorie di terzo livello)*

.

3.1.9 Altre attività specificate legate alla preparazione di alimenti

3.2 PULIZIA E RIORDINO DELLA CASA

.

. *(ALTRE CATEGORIE DI SECONDO LIVELLO)*

.

3.9 CURA E AIUTI AD ADULTI DELLA FAMIGLIA

La raccolta dei dati sull'uso del tempo è finalizzata alla misurazione della durata delle attività, ma soprattutto ad una corretta caratterizzazione delle stesse in base al contesto, alle modalità con cui si succedono, agli scopi, al luogo e al momento del giorno, della settimana, della stagione in cui si svolgono. Proprio per la molteplicità delle dimensioni a cui la descrizione delle attività riportate sui diari rimanda, la fase di codifica è il "perno" della realizzazione di tale indagine. Essa rappresenta la chiave per estrapolare un codice sintetico dalle descrizioni riportate in forma testuale libera, con l'utilizzo di espressioni eterogenee e in funzione del bagaglio linguistico e della capacità espressiva di ciascuno (Camporese e Ranaldi, 2001).

Tuttavia, come noto, l'associazione testo-codice è tutt'altro che un'operazione semplice, quanto piuttosto il risultato di un processo: 1. le espressioni verbali devono essere comprese e

valutate nel loro contesto, che è fatto di informazioni sull'individuo in oggetto e sulle caratteristiche della specifica giornata; 2. Le informazioni ridondanti non devono essere considerate, ma vanno selezionate solo quelle utili; 3. Vanno dedotte le informazioni non scritte.

I codificatori sono dunque chiamati a fare una vera e propria traduzione da un linguaggio ad un altro (quello in codici) che si pone ad un livello di complessità, ricchezza ed eterogeneità molto lontano da quello originario. Un'altro elemento di divergenza tra i due linguaggi è il diverso grado di tolleranza per l'ambiguità. Espressioni vaghe, frasi dai molteplici significati sono frequentemente usate e riconosciute nel linguaggio verbale; al contrario, il linguaggio dei codici è rigido, non è in grado di gestire le ambiguità e la necessità di interpretazione.

Tuttavia, è proprio la capacità di contestualizzare l'attività la principale competenza necessaria all'attività di codifica: per scegliere correttamente un codice occorre concentrarsi sul diario come un tutto, immaginando l'ora del giorno, la stagione dell'anno, e tenendo ben presente le caratteristiche sociodemografiche del rispondente (Camporese e Ranaldi, 2001). In sintesi due aspetti devono essere distinti. L'azione svolta e il suo obiettivo, poiché due azioni eguali, possono essere codificate in modi differenti a seconda dei loro differenti obiettivi. Solo nel momento in cui si ha questa visione di insieme è possibile concentrarsi sulle singole parti del diario e individuarne gli episodi, interpretando nel giusto modo ciascuna attività descritta (per esempio, se cucinare significa cura della propria famiglia, aiuto informale ad altre famiglie, lavoro volontario, etc.)

3. Le principali novità nella progettazione del sistema di codifica dell' indagine 2002-2003: registrazione delle informazioni testuali e codifica assistita

Uno dei primi problemi che ci si è posti in fase di riprogettazione dell'indagine ha riguardato proprio le strategie migliori di applicazione della classificazione, in termini di strumenti tecnici e di personale a cui affidare questo delicato compito.

Nella precedente indagine svolta nel 1988 e nella pilota del 1996, l'esperienza della codifica ha fatto sorgere non pochi problemi ed ha richiesto l'individuazione di strategie alternative e soprattutto innovative rispetto al passato. Per la precisione, nel corso della precedente indagine "Uso del tempo" (1988) è stata adottata la codifica manuale, effettuata dai rilevatori sulla base di apposite liste di codici relativi a "tipo di attività", "luogo di svolgimento" e "persone presenti". Tale lavoro non sempre è stato condotto in maniera ottimale, al punto che si è reso necessario un impegnativo processo di revisione che quasi sempre ha comportato la rilettura dei diari originali, per verificare l'adeguatezza dei codici e adottare gli opportuni interventi correttivi. In pratica, i revisori, per eliminare tutti gli errori e le omissioni accumulate in sede di rilevazione, hanno effettuato un lavoro simile a quello già svolto dai codificatori, cioè la lettura sul materiale cartaceo dell'attività descritta per esteso, la ricerca del codice, il confronto con il codice registrato e l'eventuale correzione.

Anche nel corso dell'indagine pilota del 1996 il processo di codifica fu svolto manualmente, questa volta ad opera di codificatori appositamente addestrati. Pure questa esperienza ha rivelato molti limiti, soprattutto in assenza di un'adeguata e costante supervisione dell'attività di codifica.

La codifica manuale inoltre richiede un elevato investimento di risorse umane e di tempo (si tenga presente la dimensione campionaria), ed accresce i rischi connessi alla discrezionalità del codificatore, poiché limita le possibilità di monitoraggio e di interventi in corso d'opera, consentiti invece da un sistema di codifica informatizzato e gestito in rete.

La complessità del sistema di codifica adottato ed in particolare la necessità di interpretare e contestualizzare le attività, prima di procedere all'assegnazione del codice corretto hanno reso poco indicato anche il ricorso ad un software di codifica automatica. Tale software infatti si presta alla codifica di informazioni (come la professione, l'elenco dei comuni, etc.) per le quali è possibile l'attribuzione univoca di un codice. Il sistema di classificazione adottato cioè prevede un unico codice per ciascuna delle voci/espressioni che tale informazione può presentare ed è unica la variabile da tenere in considerazione. Non è questo il caso del sistema di codifica utilizzato nel corso dell'indagine "Uso del tempo", visto che identiche attività possono essere codificate in maniera diversa, in base all'orario in cui vengono svolte, alle finalità, ai destinatari, etc., dunque le variabili che concorrono alla individuazione del codice corretto sono molteplici.

La crucialità del processo di codifica nell'indagine "Uso del tempo" e l'esperienza accumulata nel corso delle precedenti indagini hanno indotto l'Istituto a ricercare strategie di codifica più efficaci di quelle già sperimentate. In particolare, l'obiettivo che ci si è posti è stato quello di ridurre fortemente la discrezionalità dei codificatori e dunque il rischio di commettere errori in questa delicata fase dell'indagine, cercando contemporaneamente di contenere i tempi di realizzazione del lavoro e di rendere possibile un efficace monitoraggio in corso d'opera.

Sulla base di tali considerazioni si è deciso di implementare un sistema di codifica assistita utilizzando il software Blaise. Tale strategia, pur fornendo uno strumento elettronico che facilita, velocizza e supporta l'attività umana, lascia comunque al centro dell'attività la capacità critica del codificatore.

La fattibilità di tale scelta dipendeva tuttavia da due condizioni: 1) la possibilità di prevedere la registrazione, per la prima volta nella storia di questa indagine, di tutte le informazioni testuali riportate nei diari e, dunque, la costruzione di un archivio testuale per contenuto e dimensione mai acquisito dall'Istituto Nazionale di Statistica, ma con elevati costi di acquisizione e gestione; e 2) la possibilità di disporre di informazioni testuali che consentissero di predisporre il dizionario necessario per implementare il sistema di codifica assistita.

Alla prima esigenza l'Istituto ha risposto positivamente ed ex post è possibile dire che l'investimento fatto in questa direzione rende giustizia al grosso sforzo fatto dalle famiglie italiane per rispondere all'impegno richiesto dalla compilazione dei diari e che risulta molto chiaro dalla lettura delle descrizioni riportate, in molti casi vere e proprie storie di vita giornaliera. Certo non mancano i diari compilati in maniera lacunosa, poco chiara, talvolta evidentemente svogliata; ciononostante sono la gran parte i diari in cui i rispondenti hanno descritto scrupolosamente e con un dettaglio a volte maniacale lo scorrere del tempo e l'avvicendarsi delle varie attività svolte. Insomma la ricchezza di informazioni contenute nei diari è veramente impressionante (vedi tab. 1). La stessa codifica, per quanto resa necessaria dall' esigenza di pervenire a una sintesi dei dati e consentire analisi comparative a livello internazionale, non poteva esaurire le esigenze conoscitive che pure questa indagine è in grado di soddisfare. Pertanto la registrazione, oltre che dei codici anche delle informazioni testuali contenute nei diari, rappresenta una decisione estremamente importante, poiché apre la strada ad un'infinità di analisi di tipo innovativo: si pensi non solo all'applicazione di tecniche di text mining ma anche agli studi qualitativi, che senza tale archivio informatizzato sarebbe stato quanto meno enormemente più complesso realizzare.

Per quanto riguarda la seconda condizione, si è deciso di recuperare i diari cartacei raccolti nel corso della pilota del 1996 e di predisporre la registrazione delle informazioni testuali in esso contenute. Sono stati pertanto registrati oltre 500 diari relativi a circa 200 famiglie.

Tale opportunità ha presto rivelato anche altre positive conseguenze, poiché il recupero e il trattamento di tali dati ha messo in luce alcune attività che per rilevanza sociale necessitavano di essere estrapolate dal codice generico nel quale ricadevano per essere individuate con codici specificamente ad esse attribuite (per es. il vedere cartoni animati rischiava di confondersi con il vedere genericamente la TV).

Tabola 1. Numero di stringhe raccolte tramite i diari giornalieri per tipo di informazione

	Attività principale	Attività contemp.	Luogo	Totale
N. stringhe	1.458.312	412.609	1.220.961	3.091.882
di cui diverse	240.480	58.057	31.769	312.221
N. occorrenze	4.991.028	1.477.765	2.909.917	9.378.710
di cui diverse	29.494	12.920	9.284	37.764

Fonte: Istat, Indagine Multiscopo "Uso del tempo" 2002-03.

4. Dalla ricchezza del testo all'integrazione del sistema di codici

Come accennato è stato necessario integrare il sistema di codici proposto da Eurostat con altri codici in grado di individuare attività socialmente rilevanti. Alcuni di questi erano stati già introdotti nel corso dell'indagine pilota 1996, altri sono stati introdotti ed utilizzati per la prima volta nell'indagine definitiva.

Per esempio, sono stati previsti codici più adatti alla classificazione delle attività dei bambini, la cui vita quotidiana e la cui percezione del tempo presentano delle grosse peculiarità rispetto alle esperienze degli adulti. Le loro attività hanno caratteristiche e specificità che le rendono più difficili da incasellare in un elenco di codici, pensato per sintetizzare principalmente le attività quotidiane degli adulti.

Infatti, il sistema originario di codici si presenta a tale proposito molto lacunoso, anche perché l'Italia è uno dei pochi Paesi al mondo a prevedere la compilazione dei diari per le persone al di sotto dei 10 anni (soglia minima di età suggerita da Eurostat).

I dati dell'indagine pilota sono stati una miniera preziosa di informazioni e di descrizioni indicative del modo di esprimersi dei bambini. Sulla loro base è stato pertanto stilato un elenco aggiuntivo di attività e di codici corrispondenti che è andato ad integrare il sistema di classificazione iniziale.

Tuttavia, oltre ad integrare il sistema di classificazione adottato con ulteriori codici attività, si è ritenuto importante alla luce delle informazioni testuali contenute nei diari dell'indagine pilota prevedere una serie di codici cosiddetti ausiliari (Istat, 2002a), che supportassero le successive analisi tematiche e di qualità, portando alla luce alcune caratteristiche tipiche della descrizione che gli individui fanno del loro tempo e che sarebbero andate perse nella traduzione in codici del linguaggio comune.

In particolare, alcuni di questi codici sono un prezioso supporto al complesso processo di controllo della qualità dei diari e di eventuali correzioni perché consentono di:

- individuare le contraddizioni tra le varie informazioni presenti nei diari (tra attività principale e contemporanea, tra attività e luogo, attività e persone presenti, etc.);
- individuare sequenze non ben descritte;

- individuare diari particolarmente lacunosi di informazione e con sequenze illogiche.

Altri codici ausiliari invece consentono di:

- individuare attività di durata chiaramente inferiore alla durata minima dell'intervallo (eventi puntuali);
- individuare la concentrazione di attività all'interno di uno stesso intervallo temporale (attività multiple);
- individuare informazioni aggiuntive fornite dai rispondenti utili a specifiche analisi di qualità e che sfuggirebbero alla normale attività di codifica.

Attraverso tali codici ausiliari potranno essere individuati più facilmente sia le attività più frequentemente connesse a certe particolari tipologie di compilazione dei diari, sia - cosa non meno interessante – le modalità linguistiche ad essi connesse.

Infine, non è stata dettagliata in codici, ma sarà un interessantissimo oggetto di analisi testuale invece la presenza di molteplici espressioni che, non rimandando ad una attività (ovvero ad un uso oggettivo del tempo), ricadono in un codice generico. Si tratta di espressioni che descrivono molto frequentemente le condizioni di contesto, uno stato d'animo, ma non un'azione. In altri termini pur non essendo stato chiesto agli intervistati di fornire tale tipo di dettaglio, in numerosi casi ciò è stato fatto. Sarà pertanto interessante cercare di capire quanto la necessità di connotare "emotivamente" l'uso oggettivo del tempo possa consentire l'identificazione di specifici gruppi sociali, o ancora meglio richiedere una riflessione metodologica in vista di una riprogettazione dell'indagine stessa.

5. Gli strumenti della codifica assistita

Concretamente la codifica delle informazioni testuali è avvenuta sfruttando le possibilità offerte dal software Blaise, che mette a disposizione un modulo per la codifica assistita basato su due tipologie di ricerca, l'albero dei codici (vedi fig.2) e il dizionario (vedi fig.3), che possono essere utilizzate singolarmente e/o congiuntamente (Istat, 2002b). Le due modalità di ricerca necessitano della preparazione di due diversi files: uno contenente il sistema dei codici della classificazione, a partire dal quale Blaise organizza una struttura gerarchica ad albero e, l'altro, un elenco delle frasi-attività associate ai codici corrispondenti.

Con la registrazione delle informazioni testuali (attività principale, attività contemporanea) dei diari dell'indagine pilota si è ottenuto un archivio di frasi-attività, sulla cui base è stato costruito il dizionario da fornire al software come ulteriore e privilegiato ambito linguistico al cui interno effettuare la ricerca del codice corretto.

In effetti, mediante il dizionario si cerca di avvicinare la classificazione al linguaggio comune, utilizzato dai rispondenti e di offrire ai codificatori uno strumento che, oltre a facilitare e velocizzare l'intero processo di codifica, riduca le incertezze di attribuzione dei codici e conseguentemente la discrezionalità del loro operato.

La ricerca sul dizionario della stringa di un'attività in lettura viene effettuata dal software Blaise mediante la sua scomposizione in sottogruppi di tre lettere (detti trigrammi) e il successivo confronto tra questo insieme di trigrammi e quelli presenti nel dizionario, in cui tutte le stringhe sono state preventivamente trasformate in altrettanti insiemi di trigrammi. Il risultato di tale ricerca consiste nel mostrare al codificatore le stringhe del dizionario con il maggior numero di trigrammi in comune con la stringa in lettura (vedi fig. 3).

Questo tipo di ricerca, da un lato, ha consentito di ridurre il numero di stringhe da inserire nel dizionario; poiché, ad esempio, inserendo nel dizionario la stringa "dormire" tutte le stringhe che contengono le tre lettere "dor" (come ho dormito, sto dormendo, dormivo ecc.) avrebbero fatto riferimento ad essa, senza la necessità di essere tutte presenti nel dizionario. Dall'altro, ha reso necessario una revisione del dizionario ottenuto dopo la pura registrazione delle frasi attività, per evitare che la presenza di articoli o verbi ausiliari potesse comportare la segnalazione da parte del software, a seguito della ricerca, di una serie di attività che avessero in comune solo questi elementi con quella da codificare: quindi per esempio, nel dizionario non doveva essere presente l'espressione "**sto** guidando **l'**auto" ma "guidare auto". La ricerca basata sul confronto tra insiemi di trigrammi ha infine comportato la necessità di inserire tutte le flessioni dei verbi irregolari con diversa radice (come "cuocere" - "cotto", "andare" - "vado", ecc.) perché non si perdessero associazioni solo a causa della diversa forma del verbo utilizzato.

Anche la classificazione ha richiesto delle modifiche (vedi esempio 1) prima di essere utilizzata all'interno del sistema; era infatti necessario che le descrizioni fossero il più esplicative possibile e che i codici più generici contenessero chiari riferimenti ai codici più dettagliati ad essi sottostanti.

Esempio 1: Modifica della classificazione EUROSTAT

CLASSIFICAZIONE EUROSTAT		***CLASSIFICAZIONE PER BLAISE***	
01	**dormire**	**01**	**dormire, stare al letto malato *(il "riposo" va codificato con 032)***
011	sonno	011	dormire
012	stare a letto malato	012	stare a letto malato

Inoltre, poiché la classificazione prevede oltre ai semplici codici e alla loro descrizione, anche numerose note che facilitano il codificatore nell'interpretazione di attività che hanno significati molteplici a seconda del contesto (vedi esempio 2) è stato necessario prevedere una collocazione di tali informazioni. Se passare ad un codifica assistita on line ne avesse comportato la perdita infatti, lo strumento ottenuto non si sarebbe rivelato del tutto efficace.

Esempio 2: Note e specifiche

3.1.3.1 AIUTARE APPARECCHIARE, SPARECCHIARE (come aiuto a un familiare convivente) 4.2.1 AIUTARE APPARECCHIARE, SPARECCHIARE (come aiuto a persone non conviventi)

Per inserire nel sistema queste importanti informazioni sono state adottate più soluzioni:

- Introdurre un apposito campo "Nota" nel dizionario, contente delle specifiche sintetiche che spieghino l'uso corretto del codice (vedi fig. 3)
- Segnalare nell'albero dei codici gerarchici altre sezioni relative ad attività analoghe (vedi fig. 2 - segnalazione sul codice 4.1)

In conclusione il sistema di help, composto dal dizionario e dall'albero dei codici, costituisce un bagaglio informativo completo, necessario allo svolgimento di una corretta attività di attribuzione dei codici.

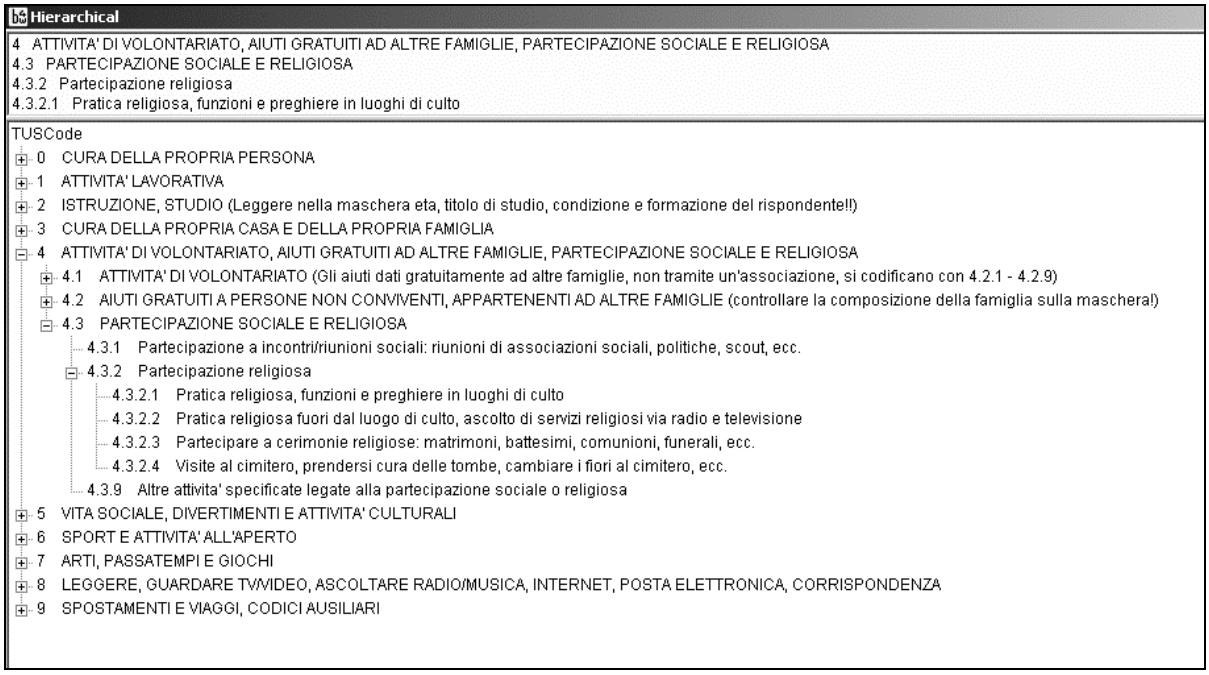


Figura 2. L'albero dei codici

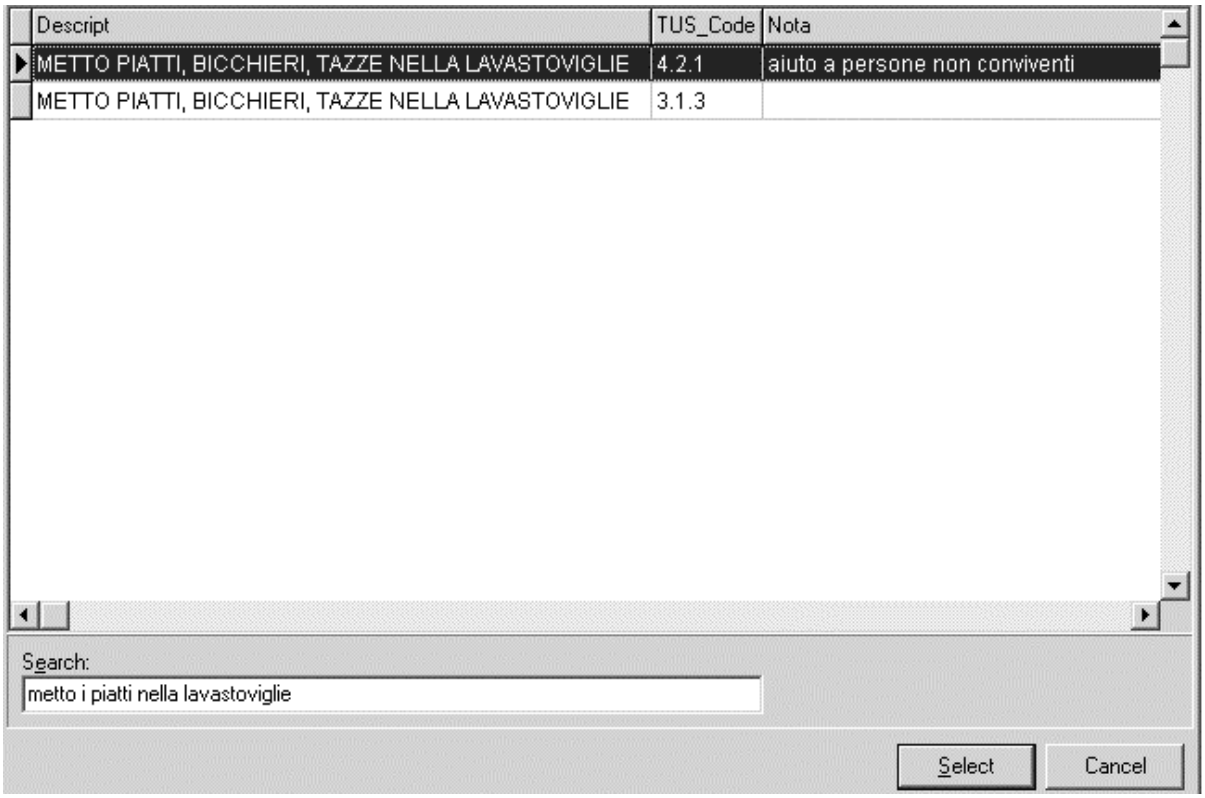


Figura 3. Il dizionario delle attività

6. La maschera Blaise per la codifica

Operativamente dunque, è stata progettata una maschera per la codifica (fig. 4-6) che ha consentito al codificatore di avere a disposizione tutte le informazioni di contesto necessarie allo svolgimento corretto del lavoro, ovvero le caratteristiche socio-anagrafiche sulla persona di cui doveva codificare la giornata, come la professione, l'età, il sesso ecc., alcune informazioni sulle caratteristiche della giornata (data, giorno della settimana), la composizione della famiglia, nonché la sequenza delle attività da registrare.

Il codificatore era tenuto a leggere attentamente le informazioni suddette e a scorrere rapidamente le attività da codificare in modo da formarsi un quadro di insieme della giornata della persona, indispensabile per la corretta traduzione delle attività in codici. Dopo aver letto le informazioni disponibili sulla persona di cui si sta codificando il diario, il codificatore poteva passare alla codifica del primo episodio registrato. Tramite il tasto F1, poteva dare avvio alla ricerca del codice sul sistema di help. La prima ricerca da tentare era quella sul dizionario, che suggerirà il codice in base alla similitudine (a gruppi di tre lettere) della frase registrata con quelle presenti nel vocabolario appositamente costruito. Se la frase era presente ed univocamente associata ad un codice, il codificatore accettava il codice e passava all'attività successiva, se invece il software suggeriva più alternative, il codificatore doveva interpretare il contesto per selezionare il codice corretto.

Nel caso in cui la frase registrata non trovasse una corrispondenza plausibile con nessun'altra presente nel dizionario, il codificatore poteva avviare la ricerca sull'albero dei codici, strutturato secondo la tradizionale forma gerarchica. In questo caso il codificatore poteva navigare l'albero in modo simile alla consultazione dell'albero delle directory in Windows, alla ricerca dell'esatto codice da attribuire all'attività registrata.

Figura 4. La maschera per la codifica assistita: le informazioni sull'individuo

Figura 5. La maschera per la codifica assistita: le informazioni sulla famiglia

Indagine Time-Use 2002: Programma per la codifica del diario

Diari Aiuto

	Inizio	Fine	Attività principale	Cod. a.p.	Attività contemporanea	Cod. a.c.	Luogo/mezzo spost.	Codl	solo	f<10	f>=10	fnc	altro	Aus.
[1]	04.00	06.40	HO DORMITO	0.1.1			A CASA	11						
[2]	06.40	07.00	MI SONO ALZATA HO FATTO COLAZIONE	0.2.1.1			A CASA	11			3			11
[3]	07.00	07.10	SONO ANDATO AL LAVORO	9.0.1.3			IN MACCHINA	53			3			
[4]	07.10	07.20	HO LAVORATO	1.1.1			AL LAVORO (CANTI	16	1					
[5]	07.20	13.40	HO LAVORATO	1.1.1			AL LAVORO (CANTI	16						
[6]	13.40	13.50	SONO TORNATO A CASA	9.0.1.3			IN MACCHINA	53			3	4		
[7]	13.50	14.30	HO PRANZATO	0.2.1.1	HO GUARDATO LA TELEVISIONE	8.2.1	A CASA	11			3	4		
[8]	14.30	15.40	MI SONO RIPOSATO	0.3.2	HO GUARDATO LA TELEVISIONE	8.2.1	A CASA	11			3	4		
[9]	15.40	16.20	HO SEGATO LA LEGNA	3.2.3			A CASA	11			3	4		
[10]	16.20	17.00	HO SISTEMATO LA LEGNA	3.2.3	HO PARLATO CON UN AMICO	5.1.2	A CASA	11			3	4		
[11]	17.00	17.30	HO SISTEMATO LA LEGNA	3.2.3			A CASA	11			3	4		
[12]	17.30	18.00	HO PULITO FUORI	3.2.2			A CASA	11			3	4		
[13]	18.00	18.30	MI SONO LAVATO	0.3.1			A CASA	11			3	4		
[14]	18.30	19.20	HO GUARDATO LA TELEVISIONE	8.2.1			A CASA	11			3	4		
[15]	19.20	19.30	HO CENATO	0.2.1.1	HO GUARDATO LA TELEVISIONE	8.2.1	A CASA	11			3	4		
[16]	19.30	20.10	HO CENATO	0.2.1.1			A CASA	11			3	4		
[17]	20.10	21.00	HO GUARDATO LA TELEVISIONE	8.2.1			A CASA	11			3	4		
[18]	21.00	21.10	SONO ANDATO AL BAR	9.0.5.1			IN MACCHINA	53	1					
[19]	21.10	21.40	HO PARLATO CON AMICI	5.1.6			AL BAR	22					5	
[20]	21.40	22.20	HO GIOCATO A CARTE	7.3.2			AL BAR	22					5	
[21]	22.20	23.00	HO PARLATO CON AMICI	5.1.6			AL BAR	22					5	
[22]	23.00	23.10	SONO TORNATO A CASA	9.0.5.1			IN MACCHINA	53	1					
[23]	23.10	04.00	HO DORMITO	0.1.1			A CASA	11						

Figura 6. La maschera per la codifica assistita: il diario

7. Riflessioni conclusive

La disponibilità dell'intero archivio dati di questa indagine oltre che rendere possibili un più ampio e innovativo ventaglio di possibilità di analisi, fornisce una base informativa importantissima ai fini di una riprogettazione metodologica delle rilevazioni sull'Uso del tempo.

Risulta chiaro infatti che a dispetto del formato del diario e delle istruzioni allegate, i rispondenti scelgono liberamente come esprimere quello che fanno e per quanto tempo.

Talvolta appare un vero è proprio gap comunicativo tra ricercatori e rispondenti, perché difficilmente questi ultimi hanno la pazienza di leggere e seguire le istruzioni. Si assiste più che ad una "comunicazione" ad una sorta di "espressione" del vissuto: nel senso che i rispondenti esprimono spesso quello che pensano ci si aspetti da loro, facendone una sintesi personale, seguendo la loro percezione e utilizzando il linguaggio che ritengono più appropriato. Infine, il tempo, non è necessariamente un percorso ripercorribile attraverso uno schema orario, la percezione rende il tempo elastico e le attività non sono sempre svolte l'una dopo l'altra. Gli individui hanno una diversa memoria della durata, una differente percezione della segmentazione del tempo in minuti e ore, una differente propensione a svolgere attività simultanee.

L'archivio testuale potrà consentire di individuare le principali tipologie di questo gap comunicativo, di analizzarlo sul piano quantitativo e fornire strumenti per migliorare l'attuale tecnica di rilevazione.

Per quanto concerne invece l'utilizzo della codifica assistita, i principali vantaggi che ne sono derivati sono sintetizzabili come segue:

- riduzione del numero di codificatori e conseguente maggiore efficacia nel processo di formazione e supervisione degli stessi (sono state previste esercitazioni pratiche e periodici de-briefing);

- monitoraggio quotidiano/settimanale, attraverso gli indicatori appositamente calcolati, dell'operato dei singoli codificatori e della loro rispondenza agli standard qualitativi prefissati;
- supervisione nella codifica di attività di difficile interpretazione;
- individuazione tempestiva di errori sistematici e conseguente adozione dei correttivi più appropriati.

In fase di progettazione si era previsto di aggiornare e incrementare periodicamente il dizionario inizialmente stilato sulla base della pilota, alla luce delle espressioni nuove e frequenti ritrovate nell'indagine in oggetto. Tuttavia, tale aggiornamento non si è reso necessario, data la soddisfacente esaustività del dizionario originario e il carattere altamente ripetitivo delle espressioni usate per descrivere la quotidianità.

Purtroppo il software Blaise non consente di calcolare in automatico le statistiche sulle diverse strategie di codifica adottate dagli operatori, ovvero in quanti casi il processo di codifica è avvenuto tramite il ricorso al dizionario, e in quanti, invece, questo si è rivelato inadeguato ed è stato necessario ricorrere all'albero gerarchico della classificazione.

Tuttavia sperimentazioni da noi effettuate su campioni di diari hanno fatto emergere che la ricerca tramite dizionario dà un esito soddisfacente ed esauriente in circa la metà delle attività da codificare.

Tanto basterebbe a confermare l'impegno necessario ad implementare un processo di codifica semiautomatica: poiché anche qualora essa non dovesse essere di aiuto nella codifica di attività rare o di significato ambiguo, comunque consente di concentrare gli sforzi interpretativi del codificatore solo su di esse.

Bibliografia

Balboni P.E. (1999). *Parole comuni culture diverse, guida alla comunicazione interculturale*. Saggi Marsilio.

Camporese R. e Ranaldi R. (2001). Time Use Activities: Translation from Sentences to Codes. Paper presentato alla Riunione scientifica del gruppo SIS (CLADAG), 5-6 luglio, Palermo.

Camporese R. e Romano M.C. (2003). Children's Time. Paper presentato alla Conferenza IATUR, 17-19 Settembre 2003, Bruxelles.

Cappadozzi T., Romano M.C., Camporese R. e Vitaletti S. (2001). Assisted coding process: an experiment in progress. Paper presentato alla Conferenza IATUR, ottobre 2001, Oslo.

Eurostat (1999). *Survey on time use: activity coding list with coding index and coding diary*. DOC E2/TUS/3.2/99.

Istat (2002a). *Uso del tempo 2002-2003. Istruzioni per la codifica delle informazioni testuali rilevate tramite il diario giornaliero*. Documento ciclostilato.

Istat (2002b). *Uso del tempo 2002-2003. Istruzioni per l'uso della maschera di codifica del diario giornaliero*. Documento ciclostilato.

Newmark P. (1988). *La traduzione: problemi e metodi*. Garzanti Ed.

Romano M.C. (2003). *Le indagini Multiscopo Istat "Uso del tempo"*. In Fraire M. (Ed.), *I Bilanci del Tempo e le indagini sull'Uso del Tempo* (Time-Budgets Studies -TBS and Time-Use Surveys-TUS), *Metodologie di rilevazione e analisi statistica dei dati sull'uso del tempo umano giornaliero*. Ed. CISU (Centro d'Informazione e Stampa Universitaria).

Shaw Valerio C. (1993). *Tips for translators*. Bulzoni Ed.

Identification des besoins en information géographique

Cristelle Roux[1], Alain Lefèvre[2]

[1]GFI Benelux S.A. – 183, route de Luxembourg – 8077 Bertrange – Luxembourg
[2]GFI Benelux S.A. – 54, Chaussée des Collines – 1300 Wavre – Belgique
cristelle.roux@gfi.be, alain.lefevre@gfi.be

Abstract

The public producers of geographic information must take into account a lot of criteria to lay down their pricing policy for the diffusion of their data. The usage of this information is one of the most difficult criteria to be quantify. GFI Benelux has developed a methodology for identifying the explicit requirements in geographical information for different potential users. This method is based on semantic and bibliometric analyses using the Tropes software in particular. This approach can be seen as the beginning of one ontology.

Résumé

Les producteurs publics d'information géographique doivent prendre en considération de nombreux critères pour définir leur politique tarifaire de diffusion de leurs données. La nature de l'usage d'une information géographique est un des critères les plus difficiles à quantifier. GFI Benelux a développé une méthodologie d'identification des besoins explicites ou non en information géographique par des usagers potentiels. Cette méthode est basée sur des analyses sémantiques et bibliométriques utilisant notamment le logiciel Tropes. Cette approche peut être le début d'une élaboration d'ontologie d'usage d'un domaine.

Mots-clés : usage, information géographique, ontologie, analyse sémantique, logiciel Tropes

1. Contexte général

1.1. *La directive*

Les administrations produisent et rassemblent, dans le cadre de leurs missions, de nombreuses données localisées ou localisables. La directive du Parlement et du Conseil européen concernant la réutilisation et l'exploitation commerciale des documents du secteur public (2003/98/CE) demande aux services publics d'établir des principes clairs de diffusion de leurs données, et d'externaliser cette diffusion autant que possible. La proposition de cette directive laisse le soin aux États membres et à ces organismes de définir leurs propres politiques tarifaires.

1.2. *La problématique : accessibilité de l'information géographique, expression du besoin*

La préoccupation actuelle des organismes publics est donc de rendre accessible à tout public les données géographiques et de définir leurs coûts tarifaires. Les études réalisées mettent régulièrement en évidence l'inadéquation des produits d'information géographique aux besoins des usagers, ce qui rend la détermination de leur valeur (et de leur prix) très complexe.

Dans le cadre de ce colloque nous limiterons la présentation de nos travaux à la notion d'accessibilité de l'information, la détermination de sa valeur économique sortant de la thématique des JADT.

Tel que le précise le projet MADAME, même si les termes « accès aux données » et « diffusion » sont souvent utilisés de manière interchangeable, une certaine différence existe entre les deux. L'accès aux données peut être envisagé comme une politique plus réactive : si un utilisateur potentiel demande à accéder à un ensemble de données, sa requête est prise en compte et l'accès peut lui être accordé moyennant une série de conditions et de frais éventuels. La diffusion de données peut être considérée comme une politique plus proactive dont l'objectif est d'encourager l'utilisation des données. Elle comprend dès lors l'accès aux données mais elle s'étend avec un effort plus positif, par exemple, à la documentation de ces données, à la fourniture d'applications, au développement de services d'assistance aux utilisateurs et elle cherche à élargir la base des utilisateurs.

D'autre part, l'accès d'utilisateurs aux données géographiques, c'est-à-dire au potentiel d'information dérivable des données géographiques, pose des problèmes au niveau de l'expression du besoin d'information géographique d'un utilisateur (Bucher, 2002). Celle-ci est très loin de la description de l'utilisation de données géographiques pouvant dériver des SIG. L'expression du besoin renvoie souvent à des concepts utilisateurs, c'est-à-dire au contexte du besoin, et non aux ressources utiles pour y répondre. Or les utilisateurs ont leurs propres représentations de l'espace géographique qui sont rarement celles utilisées par le producteur de données. Ils expriment leurs besoins avec leurs propres termes spécifiques. L'auteur (Bucher, 2002) parle ici des utilisations finales des données géographiques et non des utilisations qui consistent à construire un service de dérivation d'information pour des utilisateurs finaux potentiels.

Les applications des données géographiques sont classiquement résumées en une liste de questions élémentaires auxquelles ces applications apportent des réponses. Les deux principales questions sont :

- Où ? : La localisation. Exemple : où se trouve l'Atomium ?
- Qu'y a-t-il là ? : L'adressage. Exemple : quels sites remarquables y a-t-il dans ce pays ?

D'autres questions, classiquement identifiées dans la littérature, sont :

- Et si ? : La simulation. Exemple : Si Bruxelles n'existait pas, quels seraient les flux entre les principales villes belges ?
- Quand ? : Qu'est-ce qui a changé ? : L'analyse du changement. Exemple : quand a-t-on construit ce pont ?
- Comment ? : L'analyse spatiale. Exemple : Comment puis-je me rendre là ? Comment est structurée (géographiquement) cette ville ?

Les informations permettant de répondre à ces dernières questions nécessitent des analyses des données géographiques et des documents textuels composites résultant d'études préalables sur le même sujet ou bien sur un sujet connexe. Le SIG (Système d'Information Géographique) peut fournir les données géographiques mais il n'est pas fait pour gérer la production des documents composites.

2. Mise en évidence des usages de l'information géographique

2.1. *Choix de la méthode*

Rendre accessible une information géographique – qui découle d'une donnée géographique ou d'une combinaison de données géographiques – revient à identifier les besoins en information géographique par des usagers potentiels. Identifier ces besoins signifie également prendre en considération les caractéristiques mentionnées ci-dessous :

- La **capacité d'atteindre l'information géographique** par un utilisateur potentiel. Le niveau « d'appréhension » de l'information ne sera pas le même pour un géomètre que pour un citoyen. Il va donc dépendre d'une catégorie d'utilisateurs. Ainsi un produit fini pour un utilisateur intermédiaire sera encore inaccessible pour un citoyen.
- La **nature de l'usage** de cette information va influer sur le choix des données et des traitements nécessaires pour répondre au besoin. La nature de l'usage peut se décliner en domaine sectoriel (agriculture, urbanisme, tourisme, santé etc.) Or une difficulté majeure est l'écart entre le vocabulaire des utilisateurs et celui des bases de données géographiques. Par exemple un transporteur voulant effectuer un calcul d'itinéraire, un agriculteur voulant optimiser ses épandages, ne peuvent pas exprimer leur besoin de données géographiques à l'aide de concepts utilisés pour décrire des bases de données géographiques.

Pour mettre en évidence les besoins explicites ou non pour l'usager, nous souhaitions mettre en place une méthode ré-exploitable et objective (sans a priori) en terme de détection des besoins. Pour cela l'analyse de discours nous a semblé être la plus appropriée. Identifier un besoin par une analyse sémantique revient alors à mettre en évidence des relations sémantiques entre des acteurs et des données géographiques au sein d'un même texte ou d'un même paragraphe.

Pour soutenir ces travaux, nous avons choisi comme outil d'analyse sémantique la technologie de Tropes. Tropes est un logiciel de recherche en langage naturel et d'analyse documentaire fondé sur la compréhension du contenu qu'il a à traiter. Chaque terme significatif est inscrit dans une chaîne sémantique : un mot suffit pour retrouver tous les documents le contenant, lui et ses équivalents. Par exemple, une recherche sur « élection » retrouvera des textes contenant « voix », « urne » et « vote », même si le mot « élection » ne se trouve pas explicitement dans le texte. Tropes propose de créer des classifications de termes qui peuvent ensuite permettre d'interpréter le contenu du texte (cf. www.acetic.com).

A partir de ces objectifs et en fonction de l'outil choisi, nous avons alors décliné ces travaux selon les étapes suivantes :

- Constitution de classes :
  - De termes propres « au monde » de l'information géographique, termes géomatiques
  - D'usagers potentiels de l'information géographique
  - De termes thématiques : le domaine de l'agriculture a été choisi pour limiter notre démarche dans un premier temps.

  Remarque : pour chaque classe une validation par des experts est nécessaire. Ce processus long prendra du temps, mais un processus d'amorçage est aujourd'hui nécessaire.
- Mise en relation des termes géomatiques et des catégories d'usagers afin de mettre en évidence les usages (et donc les besoins) de l'information géographique
- Proposition d'une classification en thématique des besoins détectés

2.2. *Constitution des dictionnaires et classification des termes*

Pour constituer ces glossaires, notre démarche a été double : identification de glossaires existant sur internet ou fournis par les correspondants des administrations de la Région wallonne et identification de nouveaux termes par indexation de sites Internet.

2.2.1. *Recensement de glossaires de termes techniques en SIG et classification*

Les sources exploitées sont le dictionnaire de données de la Région wallonne (projet META-

TER), le fond de plan cartographique à grande échelle de la Région wallonne, le lot de données PLI (Plan de Localisation Informatique), le glossaire proposé par le site internet du Centre Régional de l'Information Géographique de PACA et de l'université de Laval. Lors du recensement des différents glossaires, nous constatons que les termes sont « mélangés ». On y trouve à la fois :

- Des termes très techniques purement experts : vecteurs, sections, système d'arpentage des terres, triangulation …
- Des termes techniques de la géomatique mais qui s'approchent des applications : bathymétrie, topographie, cartographie, télédétection, géomorphologie
- Des termes qui constituent des objets de cartes compréhensibles par tous : vallées, routes, forêts
- Des termes géomatiques à connotation applicative : carte maritime, pédologique, etc.

Nous proposons la classification suivante :

TERMES GEOMATIQUES
- Cartographie
  - Cartographie
  - Echelle de carte
  - Légende
  - Topographie
- Géodésie
  - Géomatique
  - Géoréférencement
  - Géoscience
  - Télédétection
- Objet géographique
  - Adresse postale
  - Information touristique
  - Objet topographique
  - Occupation du sol (Corine)
- Lots de données Métater

GEOGRAPHIE
- International
- Nation
- Pays
- Echelle géographique

2.2.2. *Recensement et classification de la liste des ACTEURS*

Les sources pour constituer la liste des acteurs proviennent du code métier (ROME) de l'agence nationale pour l'emploi française, du vocabulaire déjà existant au sein de la classe « Gens » de Tropes, du guide 2000 des administrations de la Région wallonne, de sites internet d'organismes européens SIG. Nous avons classé la liste de ces acteurs en tenant compte de l'aspect économique qui peut intervenir lors de l'usage d'une information géographique. Pour élaborer cette classification, nous nous sommes ainsi basés sur les travaux de l'ASIT-VD qui constate que, même si le marché de l'information géographique est diversifié, il en ressort deux tendances :

- **Des utilisateurs professionnels** qui ne peuvent se passer de l'information pour la réalisation de leurs activités (p.ex. services des administrations publiques, gestionnaires de réseaux, autorités locales, géomètres).
- **Des intermédiaires**. Ces derniers se composent des catégories suivantes :
  - agents qui fournissent l'information originale sans valeur ajoutée
  - éditeurs qui travaillent la donnée afin de fournir des produits et services avec valeur ajoutée
  - consultants, mandataires qui achètent et traitent l'information dans le cadre d'un mandat.

  Ces intermédiaires bénéficient en général d'un pourcentage sur la vente de la donnée originale et facturent principalement la valeur ajoutée.

– **Les autres utilisateurs** sont plus dispersés et sensibles au prix de l'information géographique car elle n'est pas indispensable aujourd'hui mais le deviendra certainement pour la réalisation de leurs activités. Par exemple, dans le cadre des lignes de vies prônées par les instances opératrices de e-Gouvernement.

Nous proposons donc la classification suivante :

FOURNISSEUR PUBLIC
- Organisation SIG

INTERMEDIAIRE
- Consultant
- Editeur

STRUCTURE
- Instance non gouvernementale

UTILISATEUR
- Instance gouvernementale de la Région Wallonne
- Autorité locale
- Citoyen
  - Qualité du citoyen
  - Groupe social
  - Profession
- Secteur privé

2.3. *Constitution de corpus pour identifier les besoins en information géographique*

La source la plus accessible en terme de gratuité et de facilité d'accès étant l'internet, nous avons constitué un corpus (ensemble de textes) à partir de plus de 500 sites internet :

– **Par acteur** : nous avons privilégié la constitution d'un sous-corpus sur les bureaux d'étude car ils nous semblaient être les interlocuteurs intermédiaires les plus à même de mettre en évidence les liens entre usagers et termes géomatiques. Nous avons également analysé des sites web d'entreprises wallonnes et d'institutions belges ;

– **Par thématique** : afin d'évaluer l'efficacité de la méthode, nous avons traité des sites internet sur la thématique de l'agriculture et plus particulièrement sur les informations que fournit l'usage de la donnée « parcelle agricole » c'est-à-dire ses caractéristiques topographiques ainsi que le type de culture.

3. Résultats

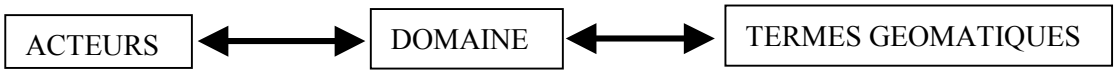
3.1. *Premiers retours*

3.1.1. *Adaptation de la méthode*

Les premiers résultats ont démontré que cette mise en relation existe au sein des textes mais qu'elle n'est pas toujours évidente car les termes géomatiques restent encore trop « experts », trop techniques : ils ne permettent pas de définir le ou les usages potentiels finaux. Nous avons alors élargi les possibilités de mise en évidence des préoccupations de positionnement d'usagers en information géographique par les 2 procédés suivants :

– Définir les données ou informations géomatiques à partir d'un langage commun encore plus simple, plus « compréhensible » par les utilisateurs. Nous avons ainsi créé une nouvelle sous-classe qui regroupe des termes génériques tels que « localisation, emplacement, site, lieu ».

– Mettre en évidence les relations entre acteurs et termes géomatiques via le domaine d'activité étudié.



La constitution et la classification d'une liste de termes propres à l'agriculture a été réalisée à

partir du vocabulaire déjà existant au sein de la classe « agriculture » de Tropes, de l'identification de nouveaux termes provenant de sites Internet en agriculture.

3.1.2. *Développement informatique*

La technologie de Tropes met en évidence les liens existants entre des termes mais elle ne permet pas de choisir des classes de termes à mettre en relation. Par conséquent, la réalisation technique de ces croisements a nécessité un développement informatique spécifique. Nous avons associé pour cela le traitement sémantique de Tropes à des méthodes de co-occurrence. Cette mise en relation des classes de termes peut être réalisée à l'échelle du texte ou du paragraphe.

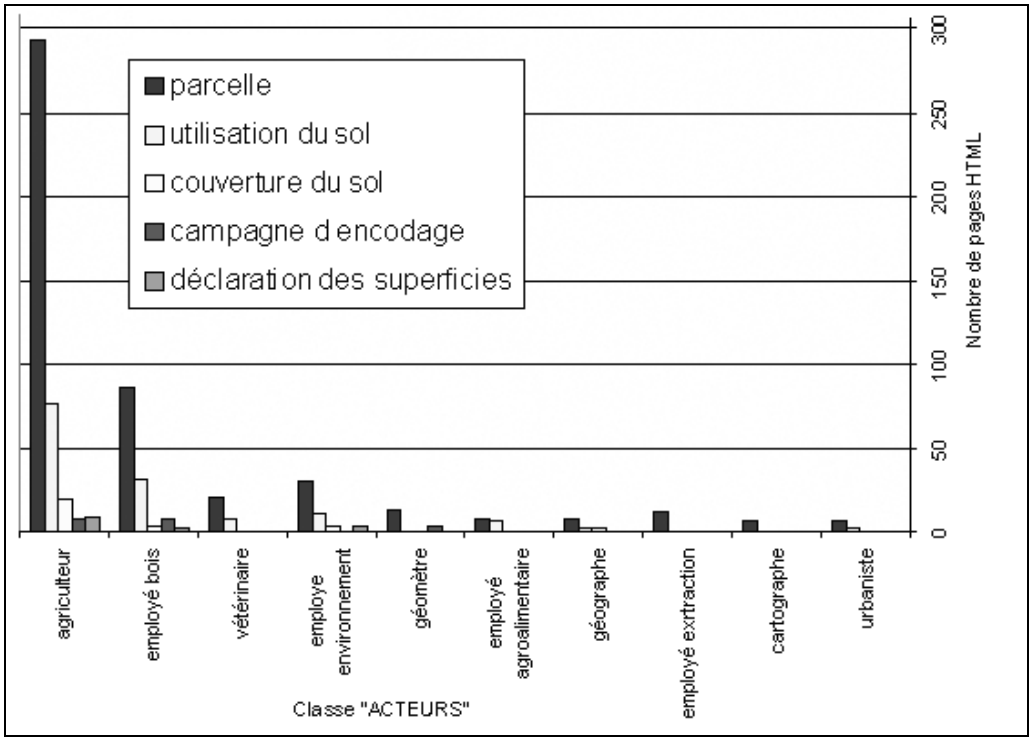


Fig. 1. Nombre de pages html contenant au sein de leur texte des termes de la classe « ACTEURS » et des termes de la classe « GEOMATIQUE »

3.2. *Mise en évidence des usages de l'information géographique*

3.2.1. *Exemples de besoins identifiés*

L'analyse sémantique du corpus a permis de mettre en évidence des usages des informations liées aux parcelles agricoles. Les textes suivants présentent quelques extraits des pages internet qui explicitent ces usages.

Réduction des largeurs et surface minimale des parcelles en jachère au bord des lacs et cours d'eau

Un règlement de la Commission européenne du 27 décembre 2000 précise que les parcelles de jachère d'un seul tenant situées le long de cours d'eau et de lacs pérennes peuvent avoir une largeur minimale de 10 mètres au lieu de 20 et une surface minimale de 10 ares au lieu de 30.

Institut Technique des Céréales et des Fourrages

Pour anticiper et adapter vos conseils techniques tout au long de la campagne, l'ITCF vous fournit des informations et des conseils de culture personnalisés (blé tendre, orge de brasserie, blé dur ?).

Repères® Conduite s'appuie sur un observatoire de parcelles suivies localement et **la mise en oeuvre de modèles agrophysiologiques** utilisant les bases de données ITCF et la météorologie locale

A chaque étape-clé de la culture, l'ITCF observe, estime, analyse les informations et rédige une synthèse comprenant : l'analyse climatique de l'année, l'évaluation des risques climatiques et leurs conséquences sur la culture, **l'état des cultures**, la prévision des stades, **la tendance du rendement** et de la teneur en protéines de la récolte.

3.2.2. *Évaluation quantitative des besoins par les catégories d'usagers*

Par les combinaisons de mise en relation entre classes présentées précédemment, on détecte les textes susceptibles de mettre en évidence des usages de l'information géographique par des catégories d'usagers. Nous traduisons ces résultats en terme quantitatif sous forme d'histogramme (fig. 1). Ces résultats apportent également une meilleure appréciation des classifications.

3.2.3. *Proposition d'une classification des usages identifiés*

Le recensement des usages démontre la multiplicité des usages de l'information liés aux parcelles agricoles. Avec l'aide d'experts du domaine, nous avons construit la classification suivante :

- **Gestion des sols** (ex : Création d'une cartographie de la végétation naturelle)
- **Gestion des territoires agricoles** (ex : Choix d'emplacement du verger)
- **Remembrement** (ex : rétablir les limites des parcelles agricoles)
- **Production-Productivité**
  - *Agriculture* **de précision** ***(ex : Optimiser les engrais de ferme dans le plan de fumure)***
  - *Les récoltes* (ex : Détection de la date de semis des betteraves)
  - *Traçabilité* (ex : Traçabilité des origines des céréales)
- **Aménagement paysager** (ex : Mesurer l'évolution entre les espaces artificialisés, les espaces agricoles et les zones « naturelles » (forêts, zones humides etc.)
- **Gestion de l'irrigation et du drainage** (ex : Aide à la gestion de grands périmètres irrigués)
- **Gestion des subventions et règlements** (ex : Obtenir prime pour mise en jachère)
- **Administration** (ex : Aide à la comptabilité, au suivi de la législation agricole, des directives de subvention, des règlements d'imposition, de l'assurance récolte.)
- **Luttes contre les maladies** (ex : Lutter contre le risque d´infiltration de la bactériose vasculaire sur les cultures de pommes de terre)
- **Risque et Sinistre** (ex : Évaluation de déclaration de sinistre)
- **Transport** (ex : Gestion de réseau de distribution pour des entreprises agricoles)
- **Impact sur environnement**
  - *Contrôle de la pollution* (ex : gérer les plans d'épandages)
  - *Les nitrates* (ex : Limiter les émissions de N2O)
  - *Gestion de l'eau* (ex : améliorer la qualité des eaux agricoles)
  - *Gestion des déchets* (ex : Droits internes)
  - *Gestion du sol* (ex : Eviter le ruissellement)
  - *Epidémiologie* (ex : l'innocuité des nitrates)
  - *Gestion des ressources naturelles* (ex : Aide à la gestion des ressources naturelles)
  - *Gestion de la faune* (ex : plans de rétablissement des espèces en péril)
- **Formation**

- *Formation*
- *Démarche* **didactique**

4. Conclusion

L'évaluation des résultats est, au stade de nos travaux, essentiellement d'ordre qualitatif. La méthode proposée permet bien de détecter des besoins en information géographique par type d'usagers potentiels. Les premiers résultats nous démontrent que la liste des termes appartenant à la classe du domaine étudié ne doit pas être exhaustive. Nous avons en effet obtenu à peu près le même nombre de pages html contenant un usage à partir d'une classe moins détaillée.

Cependant la performance de la méthode est à améliorer. Même si la combinaison des mises en relation des classes permet de ressortir les textes qui émettent des besoins, l'interprétation de ces besoins reste manuelle. L'attribution de la qualité d'un besoin tel que nous pourrions le proposer selon la classe USAGE ne pourra se faire automatiquement que par de nouveaux développements informatiques. Nous pourrions pour cela envisager d'attribuer à chaque besoin les termes qui lui sont les plus fréquemment associés. Nous identifierions ainsi une liste de termes « contextuels » de l'utilisation de la donnée géographique. Cette démarche correspond en fait aux premières étapes nécessaires pour l'élaboration d'ontologie du domaine. Toute la difficulté est par la suite de mettre en évidence de manière automatique la logique qui lie les termes pour chaque usage (ou besoin) étudié.

Références

Bucher B. (2002). *L'aide à l'accès à l'information géographique : un environnement de conception coopérative d'utilisations de données géographiques.*

CNIG. (2002). http://www.cnig.fr

COSIG.(2001). *Structure et politiques de tarification des données géographiques au sein de l'administration fédérale : étude comparative et propositions d'actions.*

CPDT. (2002). http://lepur03.geo.ulg.ac.be

Eigenheer K. *et al.* (2001). Étude pour la définition d'une politique tarifaire des données géographiques. *ASIT-VD*

EUROPEAN COMMISSION. (2002). *Directive du parlement européen et du conseil concernant la réutilisation et l'exploitation commerciale des documents du secteur publics (2002/0123 (COD))*

Heymans L. *Le fond de plan cartographique à grande échelle de la région wallonne.*

INSPIRE. (2002). *Environmental topic user needs.*

Lame G. (2000). *Acquisition de connaissance à partir de textes, vers l'élaboration d'une ontologie de droit.*

Ministère des ressources naturelles et secrétariat du conseil du Trésor du Quebec. Mise en place de l'inforoute gouvernementale : une opportunité pour favoriser la réutilisation géographique afin d'améliorer la performance de l'état et les services offerts aux citoyens et aux entreprises »

Projet MADAME (Method for Acces to Data and Metadata). (INFO 2000). Practitioners' forum on the economics of geographic information and pricing strategies.

Université de Liège LEPUR. (2002). *Le dictionnaire de données de la Région Wallonne (projet METATER).*

ProxiDocs : un outil de cartographie et de catégorisation thématique de corpus

Thibault Roy[1], Pierre Beust[1]

[1]GREYC- ISLanD CNRS UMR 6072 – Université de Caen – 14032 Caen Cedex – France
thibault.roy@etu.info.unicaen.fr, pierre.beust@info.unicaen.fr

Abstract

This paper presents a software, called ProxiDocs, which constructs a representation (a map) of the thematic structure of a whole of textual documents. ProxiDocs allows its users to realize a thematic analysis of a corpora according to his needs. These maps show thematic similarities and differences between documents which belong to a same corpora. The hypothesis tested with ProxiDocs affects the instrumentation of the meaning's thematic dimension with statistical processing, like principal components analysis and hierarchical clustering.

Résumé

Cet article présente une application dédiée à la visualisation de propriétés thématiques d'un ensemble de documents électroniques. Cette application, nommée ProxiDocs, a pour but d'assister son utilisateur dans des tâches de veille technologique en lui donnant les moyens d'une analyse thématique de corpus par rapport à ses propres centres d'intérêts. ProxiDocs produit des cartes interactives mettant en évidence les proximités et les différences thématiques des documents composant le corpus donné, ainsi que la répartition des différents thèmes utilisés dans ces documents. Les hypothèses testées dans le cadre de ProxiDocs concernent donc l'instrumentation de la dimension thématique du sens par des traitements statistiques, tels l'analyse en composantes principales et la catégorisation hiérarchique ascendante.

Mots-clés : sémantique textuelle, logiciel anthropocentré, cartographie thématique de corpus, analyse en composantes principales, catégorisation hiérarchique ascendante.

1. Introduction

Cet article prend place dans une recherche en Informatique, et plus précisément en Traitement Automatique des Langues (TAL) sur la sémantique des documents textuels. Il décrit les principes et l'intérêt d'un outil logiciel que nous avons développé au sein de notre équipe. Cet outil, appelé ProxiDocs, vise à assister son utilisateur dans des tâches de veille technologique en lui donnant les moyens d'une analyse thématique de corpus. Dans bons nombres de situations (analyse de flux documentaires, recherche d'informations, extraction d'informations …) l'appréhension de la thématique des documents ainsi que l'homogénéité thématique d'une collection de documents est une première analyse importante et souvent délicate. ProxiDocs a pour objectif de fournir à son utilisateur une aide dans ce type d'analyse en lui permettant de construire, en fonction de ses propres centres d'intérêt, des cartes thématiques de documents. En ce sens, c'est un système anthropocentré tel que le définit (Thlivitis, 1998), c'est-à-dire qu'il n'a pas une exécution guidée par des ressources propres, valables pour tout utilisateur, mais au contraire que les traitements qu'il réalise sont personnalisés car intégralement conditionnés par les besoins et les choix de l'utilisateur.

Les cartes thématiques produites avec ProxiDocs sont interactives et révèlent les proximités et les différences qui structurent un corpus supposé homogène du fait de sa provenance. Elles

permettent de mettre cette homogénéité à l'épreuve des classes lexicales choisies et définies par l'utilisateur. ProxiDocs est un outil logiciel gratuit open source, disponible avec sa documentation sur le Web (cf. http://www.info.unicaen.fr/~troy/proxidocs). C'est un logiciel d'étude au sens de (Nicolle, 1996), c'est-à-dire qu'il est conçu dans le but de vérifier des hypothèses sur les langues en les expérimentant sur du matériau textuel attesté. Les hypothèses testées dans le cadre de ProxiDocs concernent l'instrumentation de la dimension thématique du sens par des traitements statistiques. De même que d'autres outils développés au sein de notre équipe, tels que ThemeEditor (Beust, 2002) ou encore Anadia (Nicolle *et al.*, 2002), ProxiDocs fait partie d'une plate-forme de logiciels d'étude en constante évolution pour l'analyse linguistique informatisée des corpus de documents électroniques.

Nous donnerons, dans la première partie de cet article, les principes des méthodes mises en œuvre pour la construction des cartes thématiques et nous discuterons comparativement leur intérêt. La deuxième partie sera dédiée aux améliorations apportées au logiciel par l'implémentation d'une méthode de catégorisation automatique des documents d'une carte. Enfin, nous conclurons en abordant les évolutions que nous souhaitons encore apporter à l'application ProxiDocs ainsi que les pistes de recherche que le logiciel nous ouvre.

2. La cartographie thématique de corpus

2.1. *Définitions et intérêts*

À la manière du logiciel Umap (cf. http://www.umap.com), l'outil que nous proposons va chercher à mettre en évidence des similarités et des différences entre les documents d'un corpus homogène. Plus précisément, ProxiDocs va extraire les tendances thématiques des documents d'un tel corpus via une représentation graphique que nous appellerons une carte thématique. Cette représentation nous permettra de mettre en évidence les thèmes abordés dans ce corpus, mais également leur répartition et leur cohésion au sein des documents de cet ensemble. Chaque document sera représenté par un point sur la carte, la position de ce point dépendra des thèmes abordés dans le document représenté selon la propriété suivante :

> *Si deux documents abordent des thèmes similaires, alors nous supposerons que les points les représentant seront proches sur la carte.*

À l'inverse, si deux documents n'abordent pas les mêmes thèmes, alors nous pouvons faire l'hypothèse que les points les représentant seront éloignés sur la carte. Les intérêts de ProxiDocs sont alors multiples, par exemple le regroupement des documents thématiquement similaires peut aider à la création de logiciels permettant une « lecture rapide » d'un grand nombre de documents. Dans une tâche de recherche documentaire sur Internet, la cartographie thématique appliquée aux pages retournés par un moteur de recherche permettrait d'organiser les liens obtenus et de proposer des pages moins redondantes, aux thématiques différentes et se complétant.

2.2. *La construction des cartes thématiques*

2.2.1. *Les thèmes utilisés*

De la même manière que (Pichon et Sébillot, 1999), nous entendrons ici par « thèmes », les sujets abordés dans un texte. Traiter la thématique d'un texte revient donc pour nous à mettre en évidence les principaux sujets abordés dans ce dernier. Les documents électroniques que nous analysons ne sont jamais isolés et font toujours parti d'un ensemble de documents homogènes. C'est par exemple le cas des documents fournis par un moteur de recherche en réponse à une requête. Afin de construire la carte thématique d'un corpus, il sera nécessaire

de prendre en entrée les thèmes à faire émerger des documents du corpus analysé. Ces thèmes sont choisis et définis par l'utilisateur de l'application, afin de ne représenter que les thèmes et les mots pertinents de son point de vue. Techniquement, l'utilisateur va saisir ses définitions de thèmes dans un fichier XML (W3C, 2001a), compatible avec le logiciel ThemeEditor (cf. http://www.info.unicaen.fr/~beust/ThemeEditor.html). Par exemple, un utilisateur pourra définir la liste des 18 lexies (mots ou mots composés) suivantes correspondant à une définition succincte du thème « Aviation » :

Aviation : avion, avions, appareil, appareils, Airbus, A320, vol, vols, pilote, pilotes, pilotage, aéronautique, passager, passagers, Boeing, Air France, décollage, décollages.

2.2.2. *La chaîne de traitement*

Les différentes étapes de calculs de ProxiDocs s'effectuent l'une après l'autre, d'une manière chaînée. Le schéma suivant résume alors le fonctionnement de notre application :

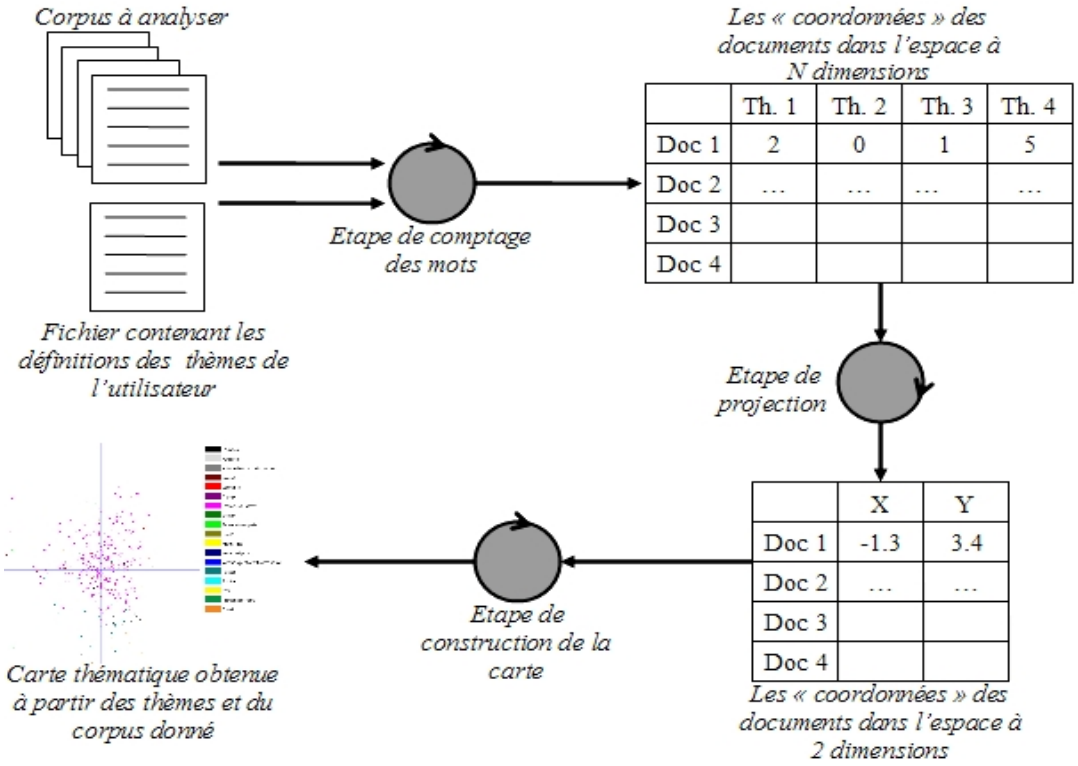


Figure 1. La chaîne de traitement de l'application ProxiDocs

Les étapes de comptage des mots et de projection sont paramétrables par l'utilisateur (les différentes possibilités seront détaillées par la suite). L'un des buts de ProxiDocs est de permettre de comparer ces possibilités. Les cartes thématiques produites en sortie de l'application sont représentées dans le format SVG (W3C, 2001b), ceci afin de permettre à l'utilisateur d'effectuer des zooms et des déplacements sur ses cartes, et de garantir la portabilité des cartes. Ce dernier pourra également visualiser les documents figurant sur la carte en cliquant sur les points les représentant. Chacun de ces points se verra attribuer une couleur correspondant au thème majoritaire dans le document qu'il représente.

2.2.3. *L'étape de comptage*

Cette étape de comptage est nécessaire pour établir une représentation vectorielle des documents du corpus étudié. Il faut compter, pour chaque document, le nombre des occurrences de mots de chaque thème. Chaque document se voit donc attribuer un vecteur contenant un nombre de données égal au nombre de thèmes utilisés. Soit *N*, un tel nombre, le vecteur décrira donc un point possédant *N* coordonnées. Soit *D* un document où figure 3 mots du thème « Bourse », 0 du thème « Météo » et 2 du thème « Sport », le vecteur obtenu sera alors : *vecteur_D* = (3, 0, 2).

La méthode précédente est appelée « absolue », du fait qu'elle ne fait pas intervenir la taille des textes lors du comptage. Une seconde méthode, appelée « relative », va chercher à déterminer pour chaque document du corpus, les proportions des mots de chaque thème par rapport au nombre de mots du document. Ainsi, pour un texte de 300 mots contenant 35 mots appartenant au thème « Sport », la méthode relative considérera que ce thème occupe (35/300*100=) 11,6% du texte. Cette seconde méthode sera particulièrement importante lorsque la taille des documents constituant le corpus varie significativement.

Une difficulté rencontrée lors du comptage des mots est qu'un mot peut appartenir à plusieurs thèmes. Ce serait par exemple le cas du mot « avocat », que l'on pourrait aussi bien affecter au thème des aliments qu'au thème de la justice. Dans un tel cas, nous avons choisi d'attribuer le mot au thème le plus représenté dans le texte. Ceci revient à prolonger le plus possible les isotopies génériques (Rastier, 1987) du texte (c'est-à-dire favoriser la redondance thématique).

2.2.4. *Les méthodes de projection*

Après application de l'une des deux méthodes de comptage précédentes, les documents sont représentés par des points possédant *N* coordonnées. La carte étant un support en deux dimensions, il va falloir projeter ces points sur ce support. Afin de réaliser une telle projection, nous nous sommes inspirés des principes généraux suivants :

i. *Déterminer les deux composantes (ou axes) « principales », c'est-à-dire les deux composantes les plus caractéristiques de l'espace de départ à N dimensions,*

ii. *Projeter l'ensemble des points de l'espace de départ à N dimensions sur les deux axes obtenus précédemment.*

Nous avons tout d'abord développé deux méthodes simples se basant sur ces principes. La première d'entre elles, appelée « méthode des deux plus grandes distances », consiste à prendre les deux axes passant par les deux couples de points les plus éloignés[1] dans l'espace de départ.

La seconde de ces méthodes simples, appelée « méthode du produit scalaire », consiste à prendre comme premier axe principal, celui reliant les deux points les plus éloignés de notre espace de départ. Le second axe sera celui passant par deux points de ce même espace et dont le produit scalaire avec le premier axe sera le plus proche de 0. Une fois les deux principaux axes de l'espace de départ déterminés par l'une des deux méthodes précédentes, il ne restera plus qu'à projeter tous les points à *N* coordonnées sur ces axes.

[1] La distance euclidienne sera utilisée pour calculer la distance entre deux points de l'espace à *N* dimensions. Ce calcul de distance pourrait être paramétrable par l'utilisateur, les méthodes de calcul de distance que nous pourrions proposer seraient, entre autres, la distance du *Khi-2* et la distance hamiltonienne.

La dernière méthode que nous avons développée est l'analyse en composantes principales (ACP) (Bouroche et Saporta, 1980). Cette méthode étant bien connue dans le domaine de l'analyse de données, nous ne la détaillerons pas ici mais nous nous limiterons à rappeler une idée de son application à notre problématique. L'idée majeure de l'ACP est que si les thèmes sont corrélés entre eux alors ils sont partiellement redondants. Prenons un exemple simple pour illustrer cette idée. Supposons que l'utilisateur ait défini les thèmes « Equitation » et « Jeux » parmi l'ensemble de ses thèmes. Supposons de plus, que son corpus soit principalement composé d'articles concernant les courses hippiques. Il est évident dans ce cas que ces deux thèmes sont fortement corrélés vis-à-vis du corpus. Ils peuvent donc être regroupés en un seul groupe de thèmes. L'ACP va généraliser ce processus à l'ensemble des thèmes, pour ne retenir que les deux principaux groupes de thèmes. Ces deux groupes sont appelés les composantes principales, la projection des documents se fera alors sur ces deux axes.

2.3. *Interprétation des résultats*

Nous allons présenter les cartes produites avec les méthodes du produit scalaire et de l'ACP. Le corpus étudié est constitué de 594 articles économiques du journal « Le Monde » dans la période 1987-1989, totalisant 500 000 mots. Le fichier de thèmes ici utilisé est défini par un utilisateur et il décrit 17 thèmes de son choix. La méthode de comptage relative est utilisée. La carte suivante sera obtenue par la méthode du produit scalaire :

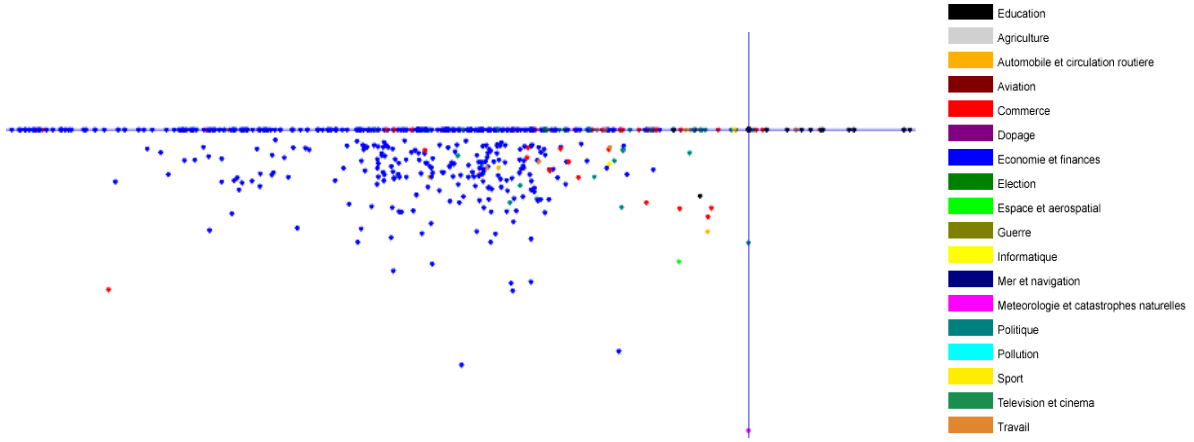


Figure 2. La carte obtenue avec la méthode du produit scalaire. Chaque point a la couleur du thème majoritaire dans le document qu'il représente. En cliquant sur l'un de ces points, on affiche le contenu du document représenté.
(version en couleur disponible sur http://www.info.unicaen.fr/~troy/proxidocs/cartes/fig2.html)

Cette carte présente les informations d'une manière particulière : les différents points sont agglomérés autour de l'axe des abscisses, et aucun point ne figure au dessus de cet axe. Ce phénomène peut s'expliquer par une relative « inefficacité » de l'axe des ordonnées choisi, puisqu'un grand nombre de points ont un projeté de valeur nul sur cet axe. Grâce à la coloration des points, nous observons tout de même une forte présence des documents abordant le thème « Economie et Finances », ceci attestant bien l'homogénéité du corpus, mais nous ne distinguons pas de groupes de documents au thème majoritaire différent.

La carte suivante a été réalisée avec les mêmes entrées que la carte précédente, la méthode utilisée est l'analyse en composantes principales :

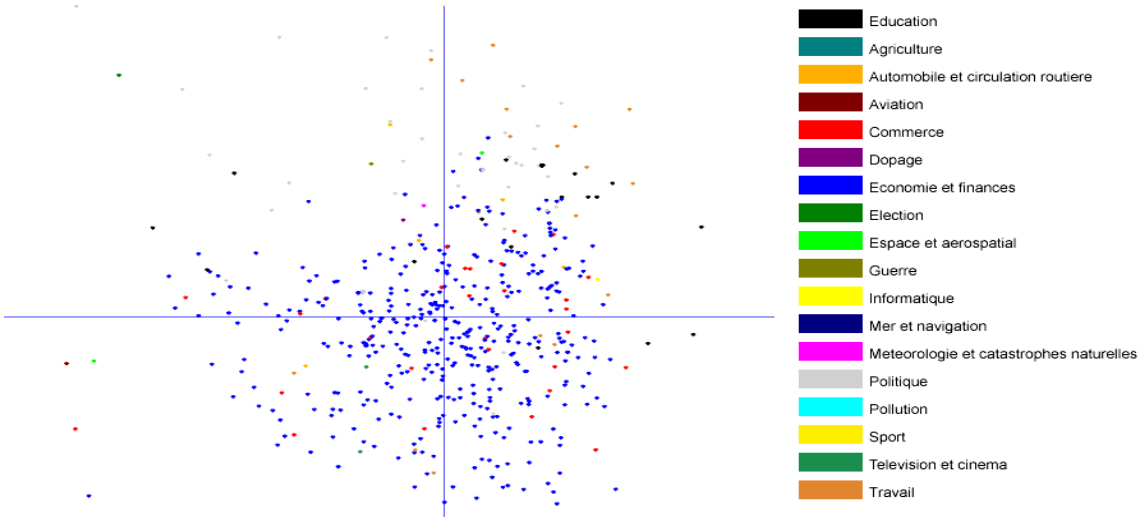


Figure 3. La carte obtenue avec la méthode de l'analyse en composantes principales (version en couleur disponible sur http://www.info.unicaen.fr/~troy/proxidocs/cartes/fig3.html)

Nous voyons que la carte met différemment en évidence les informations : les points ne sont pas agglomérés autour des axes et se répartissent mieux sur la carte. Bien évidemment, les documents de thèmes majoritaires « Economie et Finances » sont tout aussi nombreux que précédemment, mais cette fois-ci, ils sont présents dans une large zone centrée sur l'origine du repère. Des documents de thème majoritaire « Commerce » se mélangent à ces documents dans cette zone, il est donc possible de déduire un certain lien entre ces deux thèmes dans les documents du corpus. Dans la partie située en haut et à droite de la carte, des documents de thèmes majoritaires « Travail » et « Education » se mélangent. Là encore, ces deux thèmes semblent liés vis-à-vis des documents du corpus. La partie en haut et à gauche de la carte contient des documents de thème majoritaire « Politique », étant donné que ces derniers sont légèrement éloignés des autres points, nous pouvons en déduire qu'ils ne sont pas liés avec les autres documents. La carte ne laisse pas réellement paraître des groupes distincts de points, ceci attestant encore une fois l'homogénéité du corpus par rapport aux thèmes donnés.

Après quelques tests avec différents corpus et fichiers de thèmes, nous sommes en mesure de déterminer dans quelles types d'études utiliser telle ou telle méthode de projection. Sans surprise, l'ACP nous donne des cartes qui apparaissent très lisibles. Cependant, pour en extraire le plus d'informations possible une étape essentielle reste à mener : celle de l'interprétation des axes. Elle incombe à l'utilisateur (destinataire de la carte) par une analyse du contenu des documents. Les méthodes des plus grandes distances et du produit scalaire se révèlent quant à elles assez limitées pour projeter les points d'espaces de grandes dimensions. Nous obtenons cependant des résultats intéressants par ces méthodes en utilisant des espaces de « petite » taille, par exemple de 3 ou 4 dimensions.

3. Une méthode de catégorisation appliquée à la thématique de corpus

3.1. *Intérêt d'une méthode de catégorisation et intégration à la cartographie thématique*

Les cartes obtenues précédemment nous permettent de déduire des informations sur les relations thématiques existant entre les documents du corpus analysé. Malheureusement, si le corpus étudié est très homogène et les thèmes définis par l'utilisateur sont très corrélés entre eux, alors la carte obtenue ne contiendra qu'un seul groupe de points, sans mettre en évidence plusieurs groupes distincts. L'analyse d'une telle carte peut alors se révéler assez délicate.

L'intérêt d'intégrer une méthode de catégorisation est d'automatiser l'analyse manuelle de la carte en mettant en évidence sur cette carte des groupes de documents thématiquement similaires. Ainsi, l'interprétation de la carte sera alors plus simple, mais également plus fine, puisque pour chaque groupe de documents construit par cette méthode, il sera possible d'obtenir le document le plus représentatif de ce groupe.

La méthode que nous utilisons est appelée « catégorisation hiérarchique ascendante », son fonctionnement peut se résumer par les deux étapes suivantes (Bouroche et Saporta, 1980) :

i. *Parmi les n entités à classer, chercher les deux entités les plus proches. Ces deux entités sont ensuite agrégées en un nouveau groupe.*
ii. *Calculer les distances entre le nouveau groupe et les entités restantes. La configuration est alors identique à celle de l'étape i., hormis que l'on a seulement n-1 entités à classer.*

Et ainsi de suite, on cherche de nouveau les deux entités ou groupes les plus proches, que l'on agrége et ceci jusqu'à ce qu'à obtenir le nombre de groupes arbitrairement déterminé par l'utilisateur. L'application de cette méthode à notre problématique est alors assez simple, puisque nous assimilons les documents aux points d'un espace. Calculer la proximité de deux documents revient donc à calculer la distance euclidienne entre les points les représentant. Ainsi, la distance entre deux groupes de documents s'obtiendra en déterminant le centre de gravité de chacun de ces groupes, puis en calculant la distance euclidienne entre ces deux points. Le document supposé le plus représentatif d'un groupe sera alors celui étant le plus proche du centre de gravité de ce groupe.

3.2. *Intégration à la cartographie thématique*

L'étape de catégorisation va prendre en entrée les points de l'espace à deux dimensions caractérisant la carte. Cette étape de calcul va donc chercher à former des groupes de points, puis pour chacun de ces groupes, déterminer son document « caractéristique ». Nous allons présenter deux nouvelles cartes montrant les résultats de la catégorisation. Ces deux cartes seront construites à partir des mêmes entrées que les cartes précédentes, la méthode de projection utilisée sera l'ACP. Nous choisissons de faire figurer dans les deux cartes suivantes 15 groupes de documents. La première carte sera la suivante :

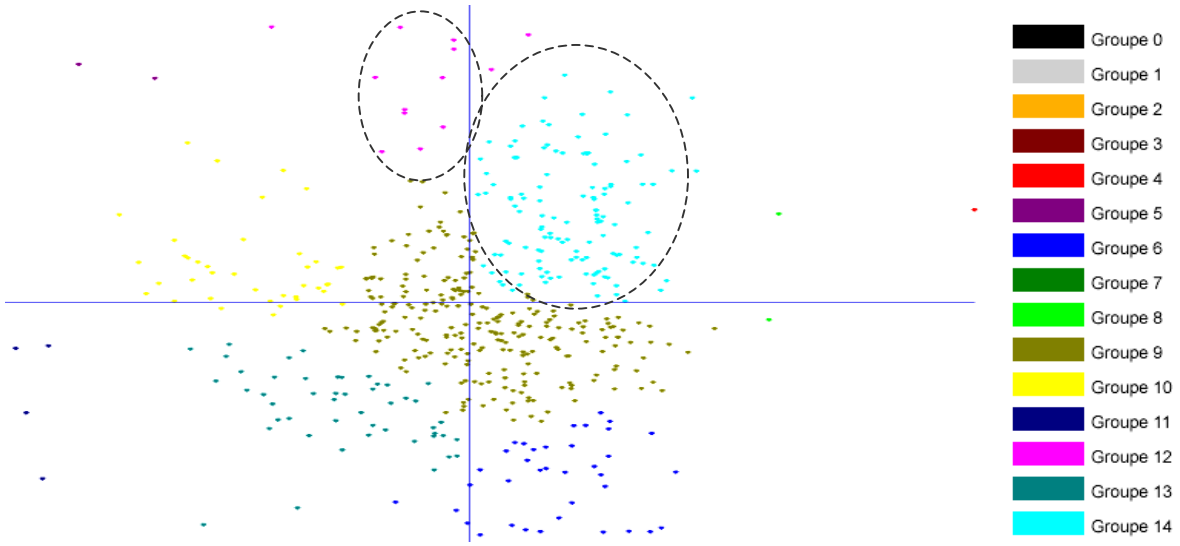


Figure 4. La carte thématique après catégorisation, chaque point est colorié suivant son groupe. Pour des raisons de lisibilité, nous avons entouré manuellement les groupes 12 et 14. (version en couleur disponible sur http://www.info.unicaen.fr/~troy/proxidocs/cartes/fig4.html)

La disposition des points sur la carte est identique à la carte de la figure 3., seule la coloration des points change. Au lieu de colorier chaque point en fonction de son thème majoritaire, les points sont colorés avec une couleur correspondant à leur groupe. De cette manière, nous voyons les groupes de documents aux thématiques similaires, ainsi que leur répartition sur la carte. En visualisant des documents appartenant à un même groupe, nous obtenons un « aperçu » de la thématique partagée par les documents de ce groupe. Ainsi, en consultant des documents du groupe 14, nous remarquons qu'ils abordent les thèmes de l'économie, de la bourse et du commerce. En légère opposition, des documents du groupe 12 abordent plus particulièrement des thèmes liés à la politique et aux actions de politiques économiques.

Afin de caractériser plus précisément chaque groupe, il est intéressant de considérer son document le plus représentatif. La carte suivante met en évidence cette propriété :

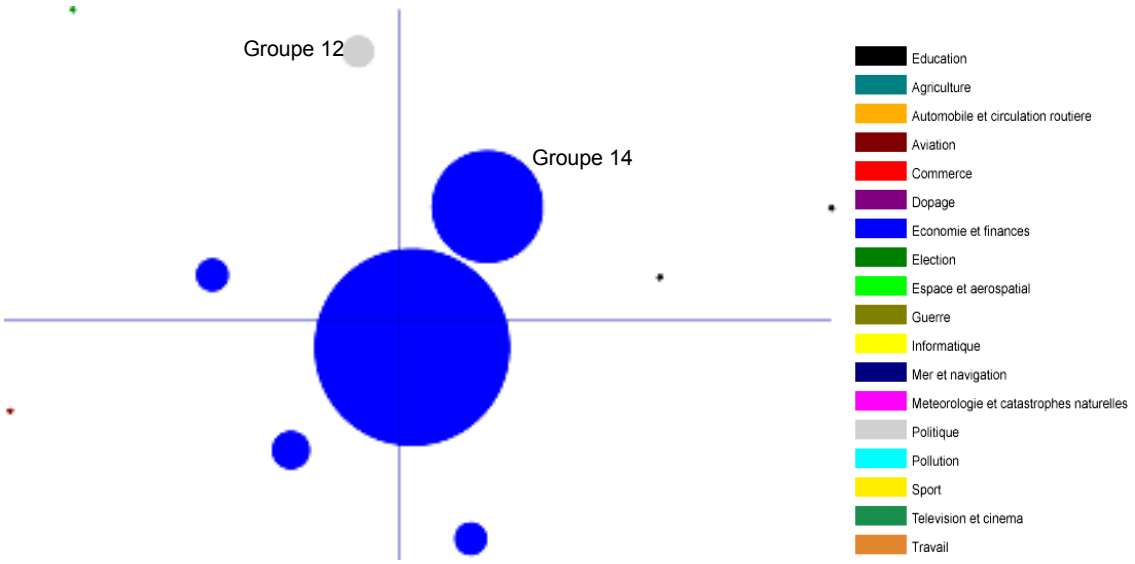


Figure 5. La carte illustrant les groupes de documents aux propriétés thématiques communes. Pour des raisons de lisibilités nous avons identifié les groupes 12 et 14. En cliquant sur un groupe, on obtient son document caractéristique.
(version en couleur disponible sur http://www.info.unicaen.fr/~troy/proxidocs/cartes/fig5.html)

Sur cette carte, un disque représente un groupe de documents. La taille de ce disque est proportionnelle au nombre de documents constituant le groupe. Chaque disque est centré sur le centre de gravité du groupe qu'il représente et il se voit attribuer une couleur correspondant au thème majoritaire dans les documents de ce groupe. Les observations que nous avons déduites de la carte précédente se voient donc facilitées puisqu'il suffit de visualiser le document le plus représentatif d'un groupe pour obtenir des informations sur la thématique des autres documents du groupe. Ainsi, les observations réalisées précédemment se confirment, puisqu'en visualisant les documents le plus caractéristiques des groupes 12 et 14, nous obtenons respectivement un document abordant une décision politique en rapport au marché boursier, et un document abordant les opérations boursières de grandes entreprises.

4. Conclusions et perspectives

Le logiciel d'étude ProxiDocs que nous venons de présenter nous a permis de vérifier notre hypothèse de départ, à savoir que l'on peut mettre à profit des méthodes statistiques pour des analyses thématiques centrées autour des besoins d'un utilisateur. De cette manière, ProxiDocs permet de vérifier l'homogénéité des corpus recueillis, d'en extraire les principales ten-

dances thématiques par rapport aux attentes de l'utilisateur, ainsi que de mettre en évidence des groupes de documents aux propriétés thématiques similaires.

Plusieurs améliorations seront envisagées dans les prochaines versions de l'application actuellement en développement. Par exemple, il serait plus facile de pouvoir définir des thèmes en entrant simplement les lemmes des lexies plutôt que des formes fléchies (ce qui demanderait à intégrer une lemmatisation à la phase de comptage des occurrences de lexies dans les documents). D'autre part, en fonction de certains types de corpus et certaines définitions de thèmes, une projection sur un espace 3D serait peut-être plus intéressante que sur une carte 2D. Pour améliorer notre application et valider plus encore nos hypothèses, il va devenir incontournable de l'expérimenter avec un plus grand nombre d'utilisateurs et des corpus d'origines diverses.

Cette expérimentation et surtout son évaluation ne sera pas sans poser problèmes car il n'est pas simplement question ici de juger (par exemple en terme de rappel et de précision) si le logiciel donne des résultats satisfaisants ou non, mais il convient plutôt d'évaluer la façon dont plusieurs utilisateurs s'approprient un même outil chacun à leur façon et en fonction de leurs buts. En somme, ce n'est pas simplement le logiciel qu'il faut évaluer mais le couple outil-utilisateur. À travers une telle expérimentation, le besoin d'autres améliorations, voire d'autres outils logiciels complémentaires, se fera sûrement sentir nous persuadant ainsi à toujours mieux instrumentaliser la dimension intertextuelle de la sémantique des langues.

Références

Beust P. (2002). Un outil de coloriage de corpus pour la représentation de thèmes. In *Actes des JADT 2002*.

Bouroche J.M. et Saporta G. (1980). *L'analyse des données*. PUF, collection Que Sais-je ?.

Nicolle A. (1996). L'expérimentation et l'intelligence artificielle. *Intellectica*, vol. (22) : 9-19.

Nicolle A. *et al.* (2002). Un analogue de la mémoire pour un agent logiciel intéractif. *Cognito*, vol. (21) : 37-66.

Pichon R. et Sébillot P. (1999). Différencier les sens des mots à l'aide du thème et du contexte de leur occurrences : une expérience. In *Actes de TALN1999* : 279-288.

Rastier T. (1987). *Sémantique interprétative*. PUF.

Thlivitis T. (1998). Sémantique interprétative Intertextuelle : assistance anthropocentrée à la compréhension des textes. Thèse d'informatique de l'Université de Rennes I.

W3C (2001a). eXtend Markup Language (XML), http://www.w3.org/XML/.

W3C (2001b). Scalable Vector Graphics (SVG) 1.0 Specification, http://www.w3.org/TR/SVG/.

Introduction à la résonance textuelle

André Salem

EA2290 SYLED – Université de la Sorbonne nouvelle – Paris 3 – France
salem@msh-paris.fr

Abstract

In order to study texts the structures of which are tightly linked such as : translations of the same text, questions and answers to them, learning material, simultaneous variations of text units through both parts of the corpus have to be considered. This paper presents a new area of textometric studies : *textual resonance* and provides an overview of the first applications of a new computing tool in various fields of textual studies.

Résumé

Lorsqu'on étudie simultanément des textes dont chacun entretient avec l'autre des rapports étroits (traductions d'un même texte, tours de paroles polémiques, données d'acquisition, etc.) il est utile de considérer les variations conjointes de différentes unités textuelles dans les deux volets du corpus. Cette étude présente les perspectives d'un nouveau domaine de la textométrie que nous appellerons la *résonance textuelle* et rend compte des premières applications à des domaines divers d'un outil informatique en cours d'achèvement.

Mots-clés : textes parallèles, statistique textuelle, textométrie.

1. Introduction

Dans le domaine des études textuelles, plusieurs situations de recherche nées dans des contextes parfois très éloignés conduisent à étudier simultanément des textes dont chacun entretient des rapports étroits avec l'autre du point de vue de sa structuration. Tel est le cas par exemple lorsqu'on considère un texte et sa traduction dans une ou plusieurs autres langues. C'est également le cas lorsqu'on considère un texte et ses différents états, versions ou brouillons, successifs ou simultanés dans une même langue ou encore les traces écrites d'une interaction verbale entre deux individus dans une situation didactique, dans une conversation ou un débat polémique.

L'étude des différents volets de ces ensembles de textes, souvent produits en simultanéité, s'enrichit de la prise en compte des relations nécessairement étroites qui lient entre elles certaines parties des deux textes. Suivant le principe qui veut que les outils informatiques doivent être pensés pour la plus grande généralité, nous proposons ici de considérer une situation très générale de cette mise en relation entre textes afin d'y adapter des outils d'exploration très généraux qui pourront ensuite être paramétrés pour des utilisations spécifiques dans des situations diverses[1].

[1] Les expériences dont nous rendons compte dans cette étude ont été réalisées à l'aide de plusieurs outils complémentaires qui appartiennent à la version actuellement disponible du logiciel Lexico3 diffusée par l'équipe SYLED. Les fonctionnalités spécifiquement dédiées à l'étude de la résonance textuelle devraient être disponibles dès la prochaine version.

2. Alignements bitextuels

Commençons par décrire une situation très générique dans laquelle deux ou plusieurs ensembles textuels sont mis en correspondance. Nous noterons ces ensembles T_1, T_2, … T_n. Nous supposons que chacun de ces ensembles est constitué par la suite des *occurrences* d'un ensemble fini de *types* (ou de formes) qui en constituent le *vocabulaire* V_i. À l'intérieur de chacun de ces ensembles, les occurrences d'items particuliers appelés *séparateurs* permettent de découper la suite des occurrences en fragments disjoints (par exemple des *phrases* et/ou des *paragraphes*). D'autres séparateurs appelées *jalons* permettent éventuellement de découper chacun des textes en un ensemble de *parties* qui peuvent correspondre à des découpages du texte que le chercheur a décidé de prendre en compte (chapitres, intertitres, tours de parole, etc.).

Dans le cas particulier de la mise en correspondance de deux textes dont l'un constitue une traduction de l'autre on appelle *alignement bitextuel* la mise en correspondance systématique de couples de fragments dont chaque élément appartient à l'un des deux textes mis en correspondance. On parle d'*alignement phrastique* lorsque les fragments mis en correspondance sont des phrases du texte, d'*alignement de paragraphes* si les unités alignées sont des paragraphes etc.[2]. On appellera *correspondances de zones* des ensembles de correspondances moins systématiques entre sous-ensembles appartenant à chacun des volets de l'ensemble bitextuel.

3. Induction et résonance textuelle

Le schéma de la résonance textuelle est illustré sur la figure 1. Cette figure représente un alignement réalisé entre deux volets d'un ensemble textuel dont on peut ignorer la nature dans un premier temps. Des fragments de texte appartenant à chacun des deux ensembles (ces fragments peuvent être des phrases, des paragraphes, des tours de parole, etc.) ont été mis en correspondance. À partir de cette mise en correspondance, toute sélection d'un sous-ensemble d'unités dans un des volets du corpus induit une sélection correspondante dans l'autre volet.

La sélection du sous-ensemble qui permet d'amorcer le processus de résonance peut être réalisée par différents moyens.

a) une sélection *topographique* constitue un sous-ensemble d'unités appartenant à l'un des deux volets exclusivement fondée sur leur localisation dans le texte (ex : les dix premiers paragraphes d'un texte qui en constituent l'introduction).
b) une sélection par *seuillage* se réalise à partir de considérations portant sur la quantité des occurrences d'un terme (ou d'une liste de termes) dans les fragments d'un des volets du bitexte (ex : ensemble des paragraphes qui contiennent des occurrences de formes commençant par le trigramme *soc*).

La sélection induite sur les fragments du second volet du corpus sert alors de point de départ à une deuxième étape qui verra la sélection des termes que leur présence (ou leur absence) rend caractéristiques pour cette dernière sélection.

[2] Les chercheurs qui travaillent dans le domaine de l'alignement automatique ont élaboré des procédures automatiques qui permettent d'aligner automatiquement les textes au niveau des phrases avec une relative efficacité (Cf. Veronis, 2000).

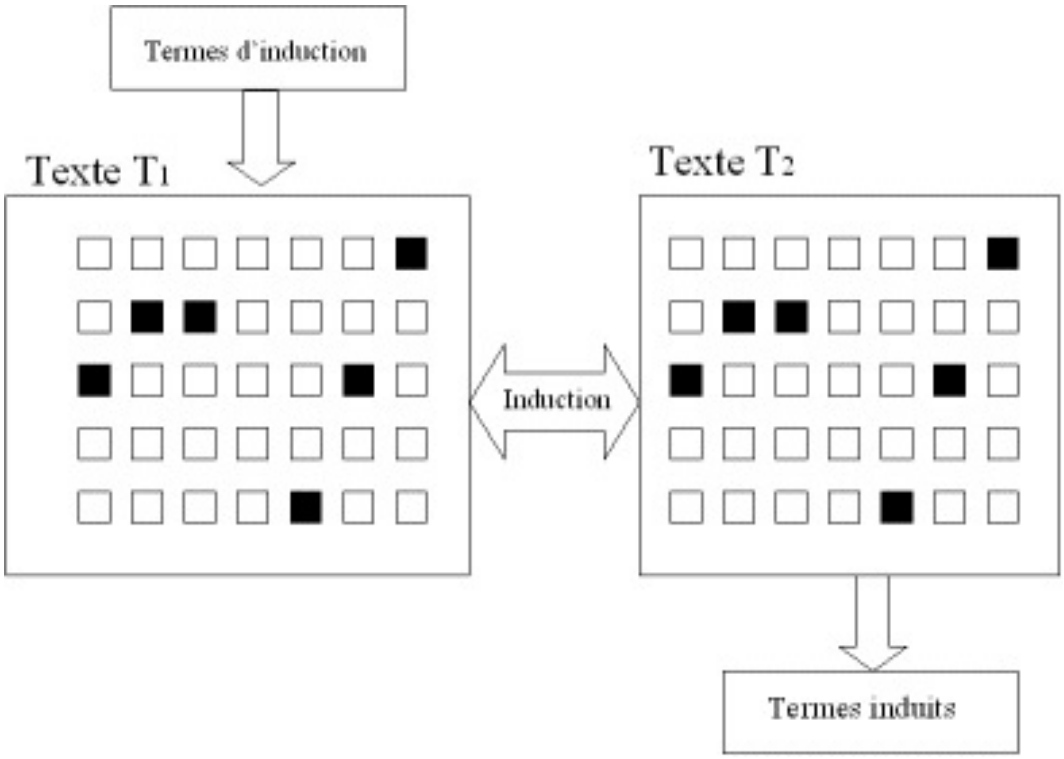


Figure 1.
Schéma général de la résonance textuelle
entre deux ensembles de textes

Guide de lecture pour la Figure 1 :

Deux ensembles textuels représentant, par exemple, deux traductions d'un même texte ont été soumis à un alignement par paragraphes. Dans chacun des volets T1 et T2 de la représentation ci-dessus chaque paragraphe du texte est représenté par un carré.

Un *terme d'induction* (par exemple un segment de texte correspondant à une locution dans le texte T1) a été présenté à un outil textométrique qui a distingué (en noir) les paragraphes du texte T1, dans lesquels ce terme trouve un nombre d'occurrences jugé important (par exemple : spécificité positive à un seuil fixé).

L'alignement en paragraphes, préalablement fourni, permet de sélectionner un sous-ensemble de paragraphes correspondants dans le second volet T2.

La dernière étape permet d'établir la liste des termes dont la fréquence est particulièrement remarquable dans le second ensemble sélectionné. Cette liste fournit des termes qui sont de bons candidats à l'appellation d'*équivalence de traduction*.

4. Applications à des domaines divers

Quelques recherches, dont certaines sont actuellement en cours, sont brièvement décrites dans ce qui suit. Ces recherches permettent d'illustrer la variété des applications possibles de la notion de résonance textuelle.

4.1. *Textes bilingues alignés*

On obtient de tels bitextes lorsqu'on met par exemple en correspondance un texte et sa traduction dans une autre langue[3]. La sélection dans l'un des volets du corpus des paragraphes dans lesquels les occurrences d'une forme, d'un segment répété ou de toute autre unité textuelle dépasse un seuil fixé provoque la sélection par *résonance* des paragraphes correspondant dans l'autre volet du corpus. La liste des unités textuelles particulièrement fréquentes dans cette seconde sélection met en évidence, dans l'immense majorité des cas des expressions qui sont des équivalences traductionnelles de la première expression.

Textes	A1 : texte juridique (en français)	A2 : sa traduction (en anglais)
Vocabulaires	V1 des formes du français	V2 des formes de l'anglais
Alignement	Alignement automatique au niveau des paragraphes entre A1 et A2	

Dans ce cas, par exemple, l'induction faite dans le volet français du corpus avec le segment répété *spéculer sur* (fr. 2 formes) renvoie le segment *speculate as to* (angl. 3 formes) qui en est la traduction. Ce type d'opération peut être réitéré en délimitant avec plus de précision la localisation des unités dont on recherche la traduction dans chacun des volets du corpus (Zimina 2000).

4.2. *Données d'acquisition*

Les textes B1 et B2 ont été recueillis dans une situation à caractère pédagogique. Un adulte converse avec un enfant à propos d'une lecture faite par ce dernier[4]. Trois séries d'entretiens ont été réalisées à plusieurs mois d'intervalle. Aux questions de l'adulte (B1) répondent les productions de l'enfant (B2). Le codage des tours de paroles permet de mettre en correspondance les questions de l'adulte et les réponses immédiatement fournies par l'enfant. Le but de la recherche est ici la mise en évidence d'interactions possibles entre le vocabulaire de l'adulte et celui de l'apprenant.

L. Sansonetti a étudié sur cette base la relation qu'on peut établir entre la fréquence des interrogations faites par l'adulte (*pourquoi ?,* etc.) et la variation ainsi que la structuration progressive dans le temps des réponses de l'enfant (*parce que*, etc.).

Textes	B1 : les productions de l'adulte	B2 : les productions de l'enfant
Vocabulaires	V1 le vocabulaire de l'adulte	V2 le vocabulaire de l'enfant
Alignement	Alignement réalisé au niveau des tours de parole de l'adulte et de l'enfant	
Partition	Une partition chronologique permet de suivre les progrès d'acquisition	

4.3. *Textes catégorisés*

À partir d'un texte C1, constitué cette fois par les réponses fournies à une question ouverte[5]

[3] Cet exemple est emprunté à des travaux menés par Maria Zimina au sein de l'équipe SYLED (Zimina, 2002). Pour une synthèse sur les problèmes du traitement des corpus parallèles, on se reportera à Veronis (2000).

[4] Cet exemple est emprunté à des travaux menés par Luiggi Sansonetti dans le domaine de l'acquisition au sein des équipes EA2290 SYLED et EA170 Calipso. (Cf. Sansonetti, dans ce même volume)

[5] Les données de cette enquête d'opinion dirigée par L. Lebart dans les années 70 ont servi de support à des exposés méthodologiques dans notre ouvrage collectif (Lebart et Salem, 1994). L'étude de l'apport des données de catégorisation a donné lieu à une publication avec B. Habert dans la revue *Traitement Automatique du Langage* (Habert et Salem, 1995).

par 2000 individus regroupés en cohorte d'âge et de diplômes équivalents, on a constitué un « texte » C2 qui contient la suite des catégories grammaticales, fournies par un outil informatisé appelé catégoriseur, qui correspondent à chacune des occurrences du texte B1. On réalise du même coup un alignement automatique qui concerne chacune des occurrences des deux textes, lesquelles se correspondent une à une.

Textes	C1 : texte des réponses à la question	C2 : catégorisation du texte C1
Vocabulaires	V1 des formes du français	V2 des catégories grammaticales
Alignement	L'alignement est alors réalisé au niveau de chaque occurrence du texte	
Partition	Les réponses sont regroupées en fonction de l'âge et du diplôme	

La sélection dans le second volet du corpus d'un patron syntaxique (ex : NOM + ADJ) permet de comprendre et d'illustrer par de nombreux exemples le fait que les personnes les plus diplômées produisent plus facilement des énoncés comportant des noms suivis d'adjectifs, les personnes moins diplômées privilégiant au contraire, la plupart du temps, les formes verbales.

4.4. *Débats télévisés*

En 1988, lors d'un face-à-face télévisé deux orateurs, F. Mitterrand et J. Chirac, prennent tour à tour la parole pour exposer contradictoirement leur point de vue[6]. On peut dans ce cas mettre en relation de résonance les tours de parole successifs (ou encore les tours de parole consécutifs, chacun étant relié à la fois à celui qui précède et à celui qui suit).

Cette mise en correspondance fait apparaître, entre autres choses, que l'emploi du mot *immigrés* par F. Mitterrand entraîne assez systématiquement celui du mot *immigration* par son contradicteur J. Chirac dans le tour de parole qui suit, et réciproquement.

F. Mitterrand :

il faut d' abord distinguer , c' est un problème qui a été vraiment exagéré et compliqué à plaisir. il y a plusieurs catégories de personnes visées par le débat actuel . il y a d - abord ceux qui ne sont pas des **immigrés** , qui sont les enfants d ' **immigrés** et qui sont nés sur notre sol . ceux - là ont vocation . ils sont français , sauf s ' ils en décident autrement à l' âge de dix - huit ans . il y a , ensuite , les naturalisés ; ce sont les **immigrés** qui désirent devenir français , là , l' administration étudie leur cas et il aboutit à reconnaître le droit à la naturalisation , selon son propre rythme . je n ' insiste pas. et puis il y a les **immigrés** _ ceux qui n ' ont pas envie de devenir français, qui veulent rester attachés à leur pays d ' origine _ de deux catégories : il y a les clandestins , et il y a ceux qui sont reconnus parce qu ' ils ont un contrat de travail et une carte de séjour . ceux qui sont clandestins , il n' y a qu ' une seule loi possible : il faut _ c ' est malheureux pour eux , mais c' est la nécessité _ il faut qu' ils rentrent chez eux et les dispositions doivent être prises , et elles ont été prises pour ceux - là , pour qu ' ils rentrent chez eux . et puis il y a ceux qui sont là avec leur contrat de travail et leur carte de séjour . est - ce qu ' il y en a trop ? ce que je sais , c' est que , dans les années qui ont précédé 1981 , il y a eu une formidable aspiration à faire venir chez nous des **immigrés** _ sans doute parce qu' on les payait moins bien que les autres , moins bien que les français , que les travailleurs français . /…/

[6] Une équipe du laboratoire de St Cloud animée à l'époque par M. Tournier a saisi ces débats télévisés afin de les soumettre à des analyses lexicométriques (Heiden et Tournier, 1998).

J. Chirac :

je voudrais répondre, moi , très clairement en m' appuyant sur mon bilan dans cette affaire ; parce que c' est très gentil de faire des promesses, mais enfin, encore faut il qu ' elles soient rendues crédibles par un bilan. s ' agissant de l ' **immigration** tout court , il faut la stopper , parce que nous n ' avons plus les moyens de donner du travail à des étrangers. aussi , naturellement , en supposant quelques souplesses naturellement, mais il faut la stopper . s ' agissant de l ' **immigration** clandestine, il faut évidemment lutter contre cette **immigration** avec beaucoup d ' énergie et reconduire les intéressés à la frontière ou les expulser. ils ont pris leurs risques en venant chez nous de façon illégale, ils sont le vivier naturel _ non pas en raison de leurs origines naturellement, mais parce que ce sont des marginaux et qui se cachent _ ils sont le vivier naturel des délinquants, voire des criminels : il faut donc les expulser. en 1981 - 82 - 83 , vous en avez régularisés 130000 : erreur capitale, car ça a été immédiatement un appel équivalent et même beaucoup plus large. nous , nous avons refoulé, en deux ans, plus de 130000 personnes, ce qui fait tout de même deux cents par jour, et je considère que ce n' est pas suffisant . nous le faisons , naturellement, en nous entourant de toutes les exigences de l' humanisme, de respect des droits de l' homme , mais c ' est une nécessité impérieuse. et puis nous devons nous protéger contre ces entrées. alors , je voudrais simplement poser une question. moi, j ' ai fait voter des lois pour la sécurité _ mais j' imagine que nous y viendrons tout à l' heure _ et contre l' **immigration** et notamment l' **immigration** clandestine, en particulier /…/

Tableau 1.
Exemple de résonance entre immigrés *et* immigration
lors du débat télévisé Mitterrand-Chirac
(présidentielles françaises 1988)

Comme on le comprend, ces deux vocables renvoient à des présentations différentes d'un même phénomène : le problème de personnes immigrées (i.e. déjà présentes sur le sol national) qu'il s'agit donc de gérer s'oppose à la présentation de *l'immigration* comme processus en cours et potentiellement menaçant.

Textes	D1 : les productions F. Mitterrand	D2 : les productions de J. Chirac
Vocabulaires	V on peut ici postuler l'existence d'un vocabulaire commun	
Alignement	Alignement réalisé au niveau des tours de parole des interlocuteurs	
Partition	Une partition chronologique permet de suivre la progression du débat	

La description de cette opposition bénéficie grandement du recours à des outils permettant l'étude directe de la résonance textuelle.

4.5. *Parentages de textes littéraires, recherche de plagiats*

L'étude des parentés/ressemblances/plagiats entre textes constitue un volet des études littéraires pour lequel les méthodes d'approche quantitative des textes sont souvent convoquées afin d'apporter un éclairage objectivant.

Textes	E1 : Un texte littéraire	E2 : Un second texte littéraire
Vocabulaires	V est le vocabulaire commun du français	
Alignement	Mises en correspondance expérimentales de zones textuelles par le chercheur	

Dans ce domaine, sauf contexte particulier, l'alignement entre différentes zones de textes dont on tente d'apprécier la proximité n'est pas donné mais posé à titre d'hypothèse sous la responsabilité du chercheur. La recherche de parentés entre les textes porte souvent sur l'agencement des unités lexicales mises en correspondance à l'intérieur des phrases. La résonance topographique constitue un outil privilégié pour le repérage de parentages.

5. Conclusion

Comme on le voit sur les quelques exemples qui précèdent, les applications du concept de *résonance textuelle* concernent un vaste ensemble de domaines de recherche liés aux approches formelles des corpus de textes.

Les applications du schéma général de la résonance textuelle à des problèmes de recherche particuliers permettent des comparaisons dont la nature peut varier d'un domaine à l'autre. Dans le cas de corpus dont l'alignement est suggéré par la nature des textes confrontés (traductions, versions différentes d'un même texte) la résonance permet de passer d'une liste de termes appartenant à l'un des volets à des termes qui lui correspondent dans le second à travers une transition fondée sur une correspondance topographique entre ces textes.

Dans le cas de confrontations verbales entre plusieurs locuteurs, la résonance peut permettre de juger de l'influence des productions de chacun des locuteurs sur celles l'autre. Dans d'autres cas (comparaison de textes littéraires etc.), l'appariement entre différentes zones appartenant aux différents volets du corpus constitue l'objet même de la recherche et doit être construit progressivement.

Les procédures informatiques qui permettent de réaliser ces expériences similaires dans des champs d'application très divers semblent par contre relever d'un modèle générique dont les paramètres pourront être modifiés par l'utilisateur en fonction de chaque projet de recherche.

Références

Habert B. et Salem A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. *Traitement automatique des langues*.

Heiden S. et Tournier M. (1998). Lexicométrie textuelle, sens et stratégie discursive http://lexico.ens-lsh.fr/biblio/slh/siad98/siad98.html

Lamalle C, Martinez W, Fleury S et Salem A. (2002). *Les dix premiers pas avec Lexico3. Outils lexicométrique*s. http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des JADT 2002.*

Lebart L. et Salem A. (1994). *Statistique textuelle.* Dunod.

Veronis J. (2000). *Parallel text processing,* Kluwer Academic publisher.

Zimina M. (2002). Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. *Lexicometrica*, numéro spécial Corpus alignés (http://www.cavi.univ-paris3.fr/lexicometrica/)

Apports de la statistique textuelle pour le repérage des reprises et reformulations dans les corpus d'interaction verbale entre un adulte et un enfant

Luiggi Sansonetti

EA 2290 SYLED, EA 170 CALIPSO – Université de la Sorbonne Nouvelle, Paris 3
19 rue des Bernardins – 75005 Paris – France
luiggi.sansonetti@noos.fr

Abstract

Besides the functions of quantification of linguistic units and comparison of use and evolution of the lexicon by author, by genre or by period, textual statistics allows to track down easily the phenomena of repetitions and reformulations in language acquisition processes in adult-child dialogue corpora.

Résumé

En plus de ses fonctions de quantification d'unités linguistiques et de comparaison d'emploi et d'évolution du lexique par auteur, par genre ou par période, la lexicométrie permet de repérer facilement les phénomènes de reprises et de reformulations dans des corpus de dialogues entre un adulte et un enfant en cours d'acquisition du langage.

Mots-clés : statistique textuelle, lexicométrie, analyse de corpus, acquisition du langage, interaction verbale adulte/enfant

1. Introduction

La linguistique de l'acquisition recherche les processus d'élaboration du fonctionnement cognitivo-langagier, en observant tout particulièrement la syntaxe chez l'enfant, à partir d'interactions verbales avec un adulte, en situation de parole spontanée. Cette approche s'appuie sur les avancées du courant interactionniste (Ochs et Schieffelin, 1995) et des travaux sur le français parlé (Blanche-Benveniste, 1997).

Dans un corpus longitudinal de dialogues entre un adulte et un enfant, nous pouvons observer des phénomènes de reprises et de reformulations de la part des deux locuteurs. Lorsque l'enfant reprend à l'adulte une construction et l'emploie dans une formulation inadéquate, et que l'adulte reformule cette construction de manière adéquate, l'adulte répond aux tâtonnements de l'enfant et lui permet alors la vérification de son hypothèse. C'est ce que Wyatt appelle « feed-back correctif » (Wyatt, 1969), désigné aujourd'hui par « interaction ajustée ou adaptée ».

Ces interactions verbales sont enregistrées puis transcrites selon des conventions de transcription propres aux objectifs de recherche. Les deux corpus utilisés dans le cadre de cet exposé ont suivi les conventions de transcription établies par Lentin et ses collaborateurs (Lentin *et al.*, 1998). Le premier corpus longitudinal, nommé « Julien », est comprend trois dialogues entre une même adulte et un même enfant, entre 5 ans 10 mois et 6 ans 4 mois. Le deuxième

corpus longitudinal, nommé « Mathilde », comprend aussi trois dialogues entre le même adulte et un autre enfant, entre 4 ans 9 mois et 4 ans 11 mois (Tissier, 2001). Pour ces interactions, l'adulte a utilisé un livre illustré *Crictor* (Ungerer, 1958) pour analyser les caractéristiques des interactions adulte-enfant dans la co-construction d'une narration.

2. Analyses « traditionnelles » et analyses informatisées

Dans le cadre de l'analyse des catégories syntaxiques d'énoncés, en particulier pour déceler l'approbation de constructions complexes par l'enfant, à partir de la transcription orthographique des dialogues enregistrés, nous recherchons dans les énoncés de l'adulte et de l'enfant des mots *(quand, puisque)*, des groupes de mots (*il faut que*, *parce que, pour que*), des constructions syntaxiques *(verbe + verbe infinitif)* ainsi que des tentatives de constructions syntaxiques chez l'enfant. Pour chaque dialogue, une grille d'analyse est remplie. Une synthèse présente un récapitulatif des éléments et des constructions trouvés, les références des énoncés contenant ces éléments, avec leurs fréquences pour chaque dialogue. L'évolution des fréquences de chaque catégorie syntaxique dans le corpus longitudinal est ensuite représentée par des courbes.

Le Projet CHILDES[1] qui présente un ensemble de programmes (CLAN) pour pratiquer diverses analyses sur des corpus de dialogues (MacWhinney, 2000), permet la recherche de formes et/ou de morphèmes, ainsi que des répétitions de formes. Cependant, il ne s'agit que d'une recherche énoncé par énoncé. Avec le logiciel de lexicométrie et de statistique textuelle Lexico3[2], il est possible de travailler énoncé par énoncé mais aussi sur l'ensemble des dialogues en même temps.

3. Repérage de « pourquoi – parce que »

Pour analyser les corpus avec Lexico3, nous avons suivi des « phases de préparation » (Sansonetti, 2003), pendant lesquelles nous avons notamment ajouté des « clés », permettant ainsi de partitionner le corpus longitudinal en dialogues ou en locuteurs (Lebart et Salem, 1994). De plus, nous avons regroupé les énoncés 50 par 50, d'abord ceux de l'adulte, puis ceux de l'enfant, et ainsi de suite. De cette façon, nous obtenons une cartographie du corpus de Julien répartie par dialogues avec sur les lignes impaires les énoncés de l'adulte, et sur les lignes paires ceux de l'enfant.

3.1. *Le corpus de Julien*

À l'aide de la fonction « Segments Répétés » de Lexico3, nous dressons la liste des suites de formes dont la fréquence est supérieure à 2. Nous recherchons, par exemple, l'emploi de la locution « parce que » et sa répartition dans le corpus longitudinal de « Julien ».

Nous constatons que le premier « parce que » est énoncé par l'enfant dans l'énoncé 20 du premier dialogue (ju=120)[3]. Cette tentative de l'enfant n'est pas reprise par l'adulte, mais par l'enfant lui-même, à l'énoncé ju=121 dans une construction syntaxique complète :

<ad=121> et pourquoi tu aimes bien ce livre ?
<ju=120> ***parce que***
<ad=122> tu aimes bien les serpents ?

[1] http://childes.psy.cmu.edu
[2] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/
[3] Selon les conventions de transcription, seuls les énoncés ne contenant que « mm » ne sont pas numérotés. C'est pourquoi, dans nos corpus informatisés, il peut y avoir un décalage dans la numérotation des énoncés.

*<ju=121> euh non mais c'est ce que je / c'est **parce que** j'avais envie*

Quand nous recherchons la motivation de ce « parce que » chez l'enfant, nous nous apercevons qu'il est toujours en réponse à un « pourquoi » de l'adulte :

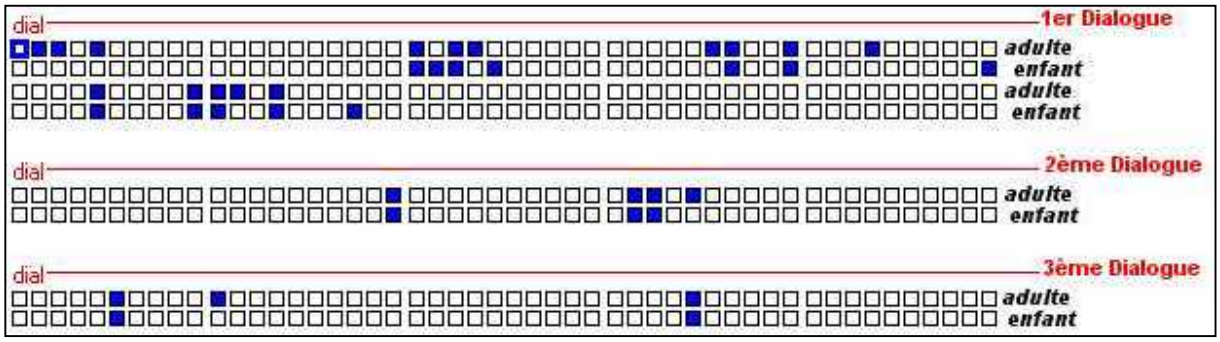


Figure n° 1. Cartographie des « pourquoi » et de « parce que » dans les corpus de Julien.

Chaque petit carré représentant un énoncé, il est facile de revenir au texte pour en vérifier le contenu. Nous avons réuni dans un même groupe de formes les « pourquoi » et « parce que ». Les carrés colorés indiquent la présence dans l'énoncé d'un élément du groupe de formes. Lorsque nous recherchons seulement les « pourquoi », nous remarquons, que l'enfant ne pose jamais ce type de question.

3.2. *Le corpus de Mathilde*

Lorsque nous procédons à la même recherche dans le corpus longitudinal de « Mathilde », nous remarquons qu'il n'y a aucune question de type « pourquoi ».

Le premier « parce que » de l'enfant, seulement énoncé dans le deuxième dialogue, est une tentative abandonnée qui n'est pas reprise immédiatement par l'adulte :

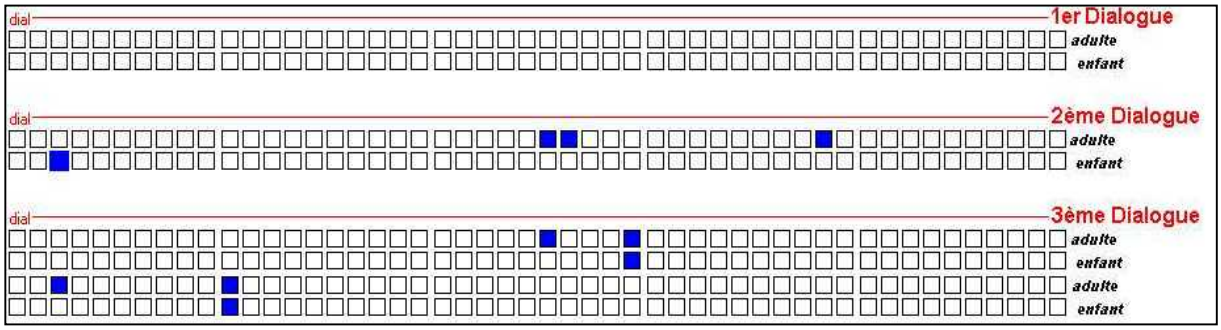


Figure n° 2. Cartographie des « parce que » dans les corpus de Mathilde

*<ma=203> elle va au zoo **parce que** le serpent trictor alors elle l'appela Crictor alors*

Dans ce troisième corpus, l'adulte énonce plusieurs constructions avec « parce que » que l'enfant ne reprend pas immédiatement. Il n'y a que trois productions de « parce que » par Mathilde dont les deux dernières, dans le troisième dialogue, sont des reprises immédiates des énoncés de l'adulte :

*<ad=329> **parce que** le mot néant commence par la lettre N*
*<ma=330> d'accord **parce que** / **parce que** heu dans / dans quoi*
[...]
*<ad=360> [...] il a un parc qui porte son nom **parce que** il a / il a / il a arrêté le cambrioleur*

<ma=360> d'accord ***parce que*** *il a arrêté le cambrioleur*

4. Repérage de catégories et de structures syntaxiques

Après avoir étiqueté les corpus « Julien » et « Mathilde » par Cordial4, nous reprenons les recherches de reprises et reformulations, non plus de formes, mais de constructions syntaxiques. Nous recherchons par exemple les constructions de type « Préposition + Verbe Infinitif » (que nous désignerons par « Prep+VInf »).

4.1. *Julien-Forme-Cat*

Dans le corpus de « Julien-Forme-Cat » (les formes sont suivies des catégories grammaticales), lorsque nous regroupons toutes les constructions « Prep+VInf » qui apparaissent au moins 2 fois dans le corpus, nous obtenons la liste suivante :

à_PREP compter_VINF	*6*
à_PREP faire_VINF	*3*
à_PREP manger_VINF	*2*
pour_PREP voir_VINF	*3*

À partir de ce groupe de formes, considéré comme une seule entité, nous recherchons les reprises de ces constructions pour en vérifier les contextes dans le corpus. Nous observons alors une production d'une construction « à+VInf » de cette structure par l'enfant après un verbe conjugué de façon non grammaticale.

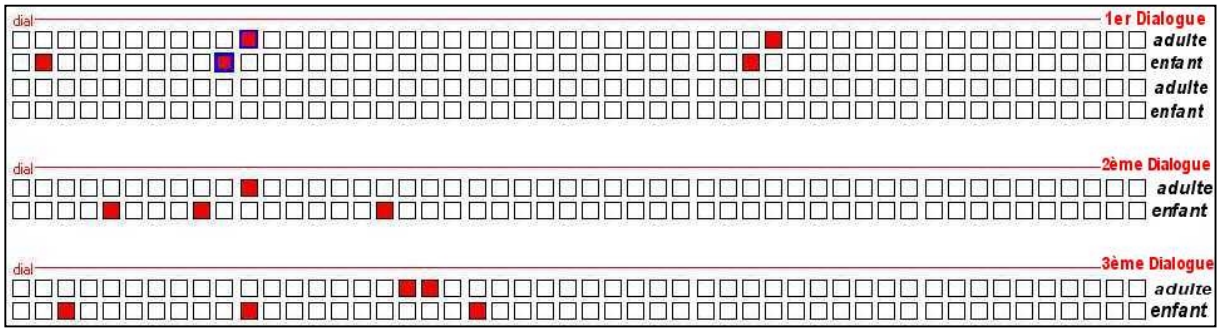


Figure n° 3. Cartographie des « Prep+VInf » dans le corpus étiqueté de Julien.

L'adulte va immédiatement reformuler l'énoncé de l'enfant avec une morphologie verbale grammaticale en utilisant le passé composé :

<ju=109> il_PPER3S apprena_NCI ***à_PREP compter_VINF***
<ad=111> donc_COO il_PREP a_VINDP3S il_PPER3S a_VINDP3S appris_VPARMS ***à_PREP compter_VINF*** *à_PREP l'_DETDFS école_NCFS aussi_ADV*

4.2. Mathilde-Forme-Cat

Dans le corpus de « Mathilde-Forme-Cat », nous remarquons que le nombre d'occurrences de la structure « Prep+VInf » est le double de celui du corpus de Julien (14 énoncés au total pour le corpus de Julien, 27 pour celui de Mathilde). Nous remarquons également avec la cartographie du corpus que les énoncés sont principalement regroupés en paires.

[4] http://synapse.com

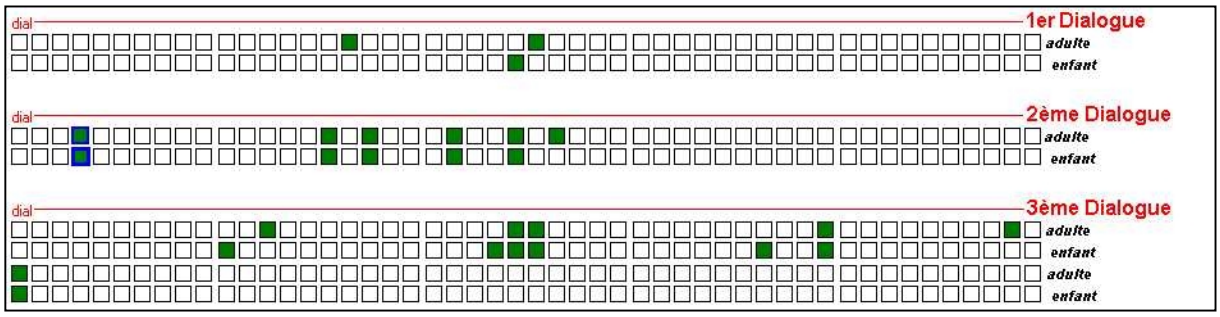


Figure n° 4. Cartographie des « Prep+VInf » dans le corpus étiqueté de Mathilde.

4.3. *Problèmes d'étiquetage automatique*

Lors de l'étiquetage de nos corpus, nous savions que les modèles de langage utilisés par Cordial sont basés sur du texte, à l'origine, écrit, et qu'il n'est pas adapté au langage parlé. Bien des études ont traité des problèmes de l'étiquetage automatique de données orales[5].

Dans un des exemples précédents, nous avons le cas de formes inconnues propres à la création de l'enfant sur le modèle des verbes du premier groupe comme « apprena ». Dans les paramétrages de l'étiqueteur, nous avons donné l'instruction de ne pas corriger la faute. L'étiqueteur l'a bien repérée mais n'a rien fait, puisque, dans notre étude qui peut être aussi lexicale ou morphologique, nous voulions conserver la forme énoncée par l'enfant. Ainsi dans le temps, il serait possible de voir si l'enfant, à force ou non de recevoir des constructions au passé simple et/ou des corrections adaptées de la part de l'adulte, passe de « apprena » ou « metta » à « apprit » et « mit ».

Une autre erreur d'étiquetage plus délicate qui s'est produite est le fait d'avoir plusieurs étiquettes ou aucune pour les mêmes formes ou mêmes locutions. Dans le corpus « Julien-Forme-Cat », nous avons *pour_PREP* et *pour qu'_SUB*. Alors que dans le corpus « Mathilde-Forme-Cat », nous avons *pour_PREP, pour qu'/que_SUB* et *pour_NCMS*. Même si pour les conjonctions de subordination, le *pour* n'est pas étiqueté isolément, ceux qui sont notés NCMS (nom commun masculin singulier) sont dus à un article précédant la préposition :

<ad=337> […] six comme les ***pour*** *les six pattes d'un insecte […]*

5. Le corpus *Julien-Mathilde*

Puisqu'il s'agit du même adulte dialoguant avec Julien puis Mathilde, nous réunissons les deux corpus en un seul, ce qui permet comparer les enfants entre eux, ainsi que l'adulte à elle-même. Nous confirmons le fait qu'il y a presque le double d'occurrences de la construction *« Prep+VInf »* chez « Mathilde » par rapport à « Julien ».

En faisant un graphique de ventilation pour une partition donnée, il est possible d'observer l'évolution de l'emploi de la forme dans les corpus, ici dans le temps. Nous obtenons le graphique suivant :

[5] Nous pouvons citer les travaux de Valli et Véronis (1999), Paroubek et Rajman (2000) et Véronis (2000).

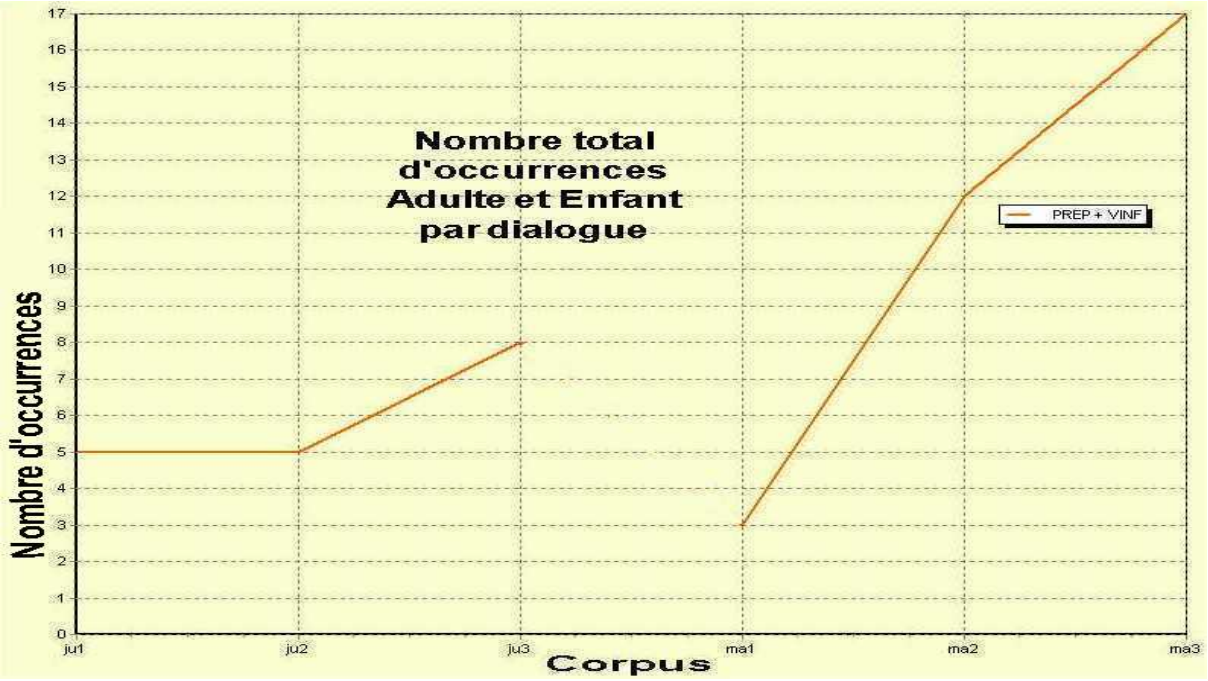


Figure n° 5. Comparaison de l'évolution de la construction « Prep+VInf » dans les corpus étiquetés de Julien et Mathilde réunis.

Nous constatons que dans les dialogues avec Mathilde, il y a de plus en plus d'occurrences de cette construction. Lorsque nous sélectionnons la partition « locuteurs », nous pouvons alors observer dans la figure suivante, dans un premier temps, que le nombre d'occurrences de cette construction chez l'adulte et l'enfant correspondant (ad1 avec ju1 et ad2 avec ma2) suit la même évolution. Nous pouvons observer, dans un deuxième temps, que Mathilde a eu plus d'offres de « Prep+VInf » que Julien de la part de l'adulte.

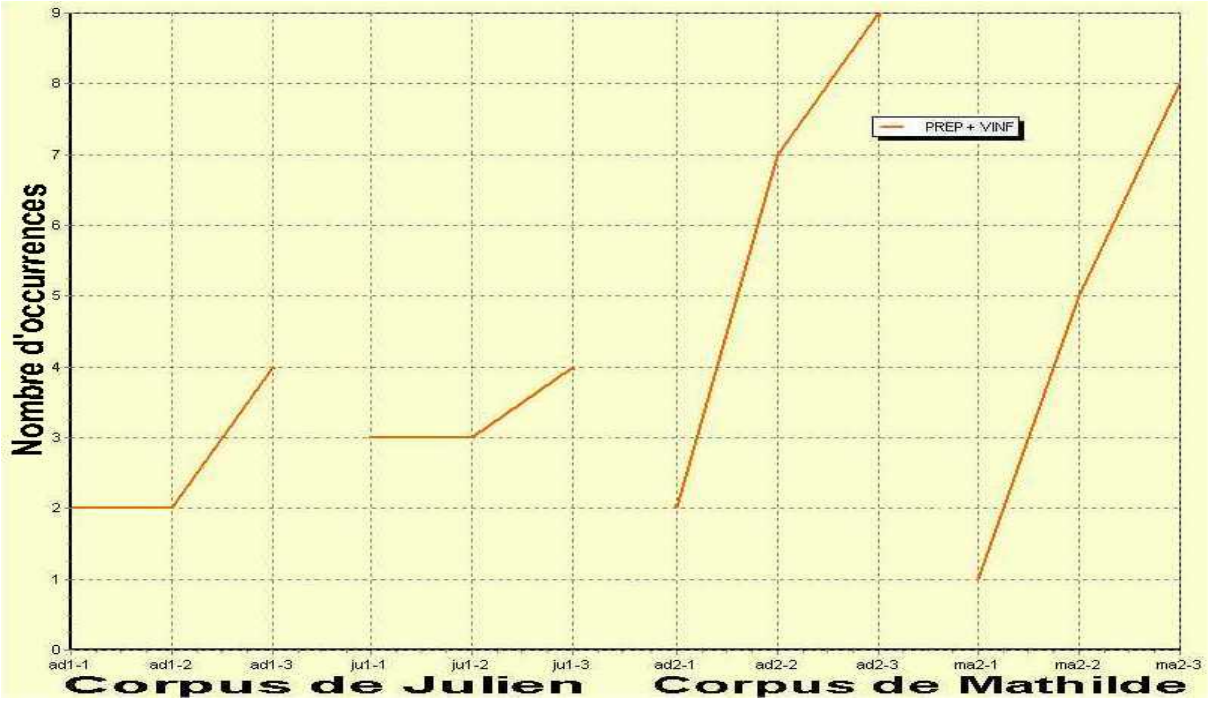


Figure n° 6. Comparaison de l'offre de la construction « Prep+VInf » de l'adulte et de sa reprise par chaque enfant dans les corpus étiquetés de Julien et Mathilde réunis.

6. Conclusion

À l'aide de logiciels de statistique textuelle et d'étiquetage automatique, il est possible d'étudier des données orales et d'observer l'évolution des emplois du vocabulaire et de la syntaxe. Il est également possible avec le logiciel Lexico3, après des phases de préparation du corpus, de visualiser rapidement les reprises et les reformulations qui sont des phénomènes pertinents dans le processus d'acquisition de la langue en situation de dialogue.

Il est montré par notre expérience que plus l'adulte propose de mots, des constructions différentes et des réutilisations dans des contextes différents, plus l'enfant essaie de reproduire puis de produire ces éléments de façon spontanée. C'est le cas de Mathilde pour les constructions du type « Prep+VInf ». Bien sûr, il faudra revenir sur les résultats des étiquetages automatiques de données orales, en nous laissant la possibilité de corriger les étiquettes avant deuxième traitement.

Références

Blanche-Benveniste C. (1997). *Approches de la langue parlée en français*. Ophrys, Collection l'essentiel français.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.

Lentin L. *et al.* (1984). *Recherches sur l'Acquisition du Langage*, t. (1). Presses de la Sorbonne Nouvelle.

Lentin L. *et al.* (1988). *Recherches sur l'Acquisition du Langage*, t. (2). Presses de la Sorbonne Nouvelle.

Lentin L. (1997). *Apprendre à penser, parler, lire, écrire*. ESF.

MacWhinney B. (2000). *The CHILDES project. Tools for analyzing talk* (2 vol.). Mahwah, Lawrence Erlbaum Associates.

Ochs E. et Schieffelin B. (1995). Language socialization and its consequences for language development. In Fletcher P. et MacWhinney B. (Eds), *The Handbook of Child Language*. Blackwell Publishers : 73-94.

Paroubek P. et Rajman M. (2000). Étiquetage morpho-syntaxique. In Pierrel J.-M. (Ed.), *Ingénierie des langues*. Hermès Sciences Publications : 131-150.

Sansonetti L. (2003). Approche lexicométrique de corpus d'interactions verbales entre un adulte et un enfant en cours d'acquisition du langage. Résultats d'expérience. Actes du Colloque « 3es Journées de la Linguistique de Corpus », Lorient, 11, 12 et 13 Septembre 2003. Presses Universitaires de Rennes (à paraître).

Tissier C. (2001). *Rôle de l'adulte dans l'interaction langagière adulte-enfant (entre 4 ans 9 mois et 6 ans 4 mois) en situation de narrations dans deux corpus longitudinaux*. Mémoire de Maîtrise, ILPGA, Paris 3 (non publié).

Ungerer T. (1958). *Crictor*. L'Ecole des Loisirs pour l'édition en langue française, 1980, réédition 2000. Collection lutin poche.

Valli A. et Véronis J. (1999). Étiquetage grammatical des corpus de parole : problèmes et perspectives. *Revue Française de Linguistique Appliquée*, vol. (IV/2) : 113-133.

Véronis J. (2000). Annotation automatique de corpus : panorama et état de la technique. In Pierrel J.-M. (Ed.), *Ingénierie des langues*. Hermès Sciences Publications : 151-171.

Wyatt G. (1969). *La relation mère – enfant et l'acquisition du langage*. Mardaga.

Hyperliens et recherche d'information sur le Web

Jacques Savoy, Yves Rasolofo

Institut interfacultaire d'informatique
Université de Neuchâtel – Pierre-à-Mazel 7 – 2000 Neuchâtel (Suisse)
Jacques.Savoy@unine.ch

Abstract

Today search engines are very useful tools for retrieving information on the Web. When submitting a query however users may not always be interested in retrieving long lists of sites; rather they may prefer only a single Web site (e.g., when looking for a given homepage, an on-line service or a named page). Or, they may be only seeking useful starting points from which to browse a given topic; thus the system should respond with a list of key resources on a given subject (topic distillation). This paper describes our search system that would respond to these two search types and how its reliance on hyperlinks enhances retrieval effectiveness. Using a relatively large Web test collection (18.1 GB), an evaluation of our strategy indicates that the presence of hyperlinks improves search precision and that our system provides performance levels superior to those of other models.

Résumé

Les moteurs de recherche s'avèrent des outils indispensables afin de retrouver de l'information sur Internet. Cependant, les usagers ne désirent pas toujours une liste de sites répondant à leurs requêtes mais souhaitent obtenir un site comme unique réponse (recherche d'une page d'accueil ou d'un service en-ligne). D'autre part, nous voulons parfois extraire une liste des ressources clés ayant trait à un domaine précis. Dans ce cas, nous ne cherchons pas toutes les pages sur un thème mais souvent des bons points de départ pour la navigation. Cet article décrit notre système de recherche capable de répondre à ces deux types d'interrogation en recourant aux hyperliens afin d'accroître la qualité des réponses. Une évaluation basée sur une collection de pages Web (18,1 GB) indiquent que la présence des hyperliens permet d'accroître la qualité du dépistage.

Mots-clés : recherche d'information, Web, recherche de service en-ligne, recherche de ressources clés.

1. Introduction

La présence de moteurs de recherche de bonne qualité a certainement été l'un des facteurs qui a permis à Internet de grandir tout en permettant un accès à un volume très important de documents. Des études récentes (Spink *et al.*, 2001) démontrent que les requêtes adressées à ces moteurs varient avec les années passant d'un Internet ludique à un Web plus commercial. Cependant, ces interrogations s'expriment toujours avec un nombre très restreint de mots. De plus, les internautes restent peu soucieux du rappel c'est-à-dire de dépister un grand nombre de pages pertinentes à leur requête. Ils attachent une grande importance à la simplicité d'accès (Nielsen, 2000) et à un temps de réponse minimal.

Actuellement, les usagers d'Internet attendent plus des moteurs de recherche. Par exemple, en réponse à la demande « cinéma Paris », le moteur de recherche devrait proposer une liste des films à l'affiche correspondant au lieu indiqué explicitement ou implicitement, tout en tenant compte de la date et de l'heure de l'émission de la requête (Watters et Amoudi, 2003). Parfois, les internautes ne désirent pas des pages relatives au thème indiqué, mais une réponse précise comme, par exemple, pour la demande « Quelle est la hauteur de la tour Eiffel ? » ou « Qu'est-ce qu'un neutron ? » (Radev *et al.*, 2002). De plus, on sait souvent que la réponse

attendue correspond à un site précis, comme lorsque l'on recherche une page personnelle, le service de réservation d'une compagnie de chemin de fer ou la page d'accueil d'une entreprise (*homepage and named page searching*). Enfin, on souhaite parfois que le moteur de recherche puisse extraire d'Internet une liste de sites proposant un contenu très pertinent ou un bon point de départ pour la navigation sur un thème donné (*topic distillation*).

Afin de répondre à ces deux derniers types d'interrogation, nous avons développé un moteur de recherche spécialement adapté à la recherche de pages précises ou capable d'extraire les ressources clés du Web sur un thème donné. La deuxième section explique en grandes lignes les principes de notre système de recherche. La troisième section décrit avec plus de détails notre stratégie de dépistage de sites spécifiques et évalue nos propositions avec l'aide d'une collection de pages du Web (1'247'753 pages pour un volume d'environ 18,1 GB). La quatrième section expose notre système d'extraction des bonnes pages de départ pour la navigation, évalue notre proposition et la compare à diverses autres stratégies.

2. Idée de base de notre stratégie de dépistage

Afin de concevoir un moteur de recherche d'information efficace pour le Web, des études précédentes ont démontré que le modèle probabiliste Okapi (Robertson *et al.*, 2000) possède une bonne qualité de réponse (Savoy et Picard, 2001). Cependant, dans ces expériences, il s'agissait de trouver tous les documents répondant à une requête. Dans le cas présent, nous ne désirons pas retourner toutes ces pages mais seulement un site correspondant au service ou contenant l'information exigée. Ainsi, face à la requête « Air France », notre système doit retourner la page permettant une réservation en-ligne auprès de cette compagnie aérienne et non pas un ensemble de documents concernant de près ou de loin cette compagnie.

Dans ce but, nous avons décidé de modifier ce modèle probabiliste afin de pouvoir tenir compte de plusieurs représentations (mais au maximum trois) des pages Web. Ainsi, nous pourrions, par exemple, générer une première indexation des pages basée uniquement sur les balises <TITLE> et <META>, une deuxième indexation sur les étiquettes <H*i*> et une troisième sur tout le texte. Lors de l'estimation du score d'une page D_i, valeur dénotée $RSV(D_i)$, nous calculons une somme pondérée du score obtenue par chaque représentation, soit :

$$RSV(D_i) = \alpha \cdot \sum_{j=1}^{m} w_{ij}^{(1)} \cdot qw_j + \beta \cdot \sum_{j=1}^{m} w_{ij}^{(2)} \cdot qw_j + \gamma \cdot \sum_{j=1}^{m} w_{ij}^{(3)} \cdot qw_j \quad (1)$$

dans laquelle $w^{(1)}_{ij}$ indique le poids attribué au terme t_j dans le document D_i selon la première représentation ($w^{(2)}_{ij}$ et $w^{(3)}_{ij}$ indique cette même pondération relative à la deuxième, respectivement, troisième représentation), et les paramètres α, β, γ indiquent l'importance accordée à chaque représentation. Enfin, le poids du terme t_j dans la requête est indiqué par qw_j et m indique le nombre de mots communs entre le document et la requête.

Habituellement en recherche d'information, il se révèle utile de permettre des appariements entre des formes variant en genre ou en nombre. Ainsi, si un article possède le terme « chevaux », il est raisonnable de penser qu'une requête contenant le mot « cheval » devrait pouvoir dépister cette page. Comme notre moteur doit posséder une haute précision (c'est-à-dire être capable de retourner uniquement des sites correspondant aux critères de la requête), nous avons renoncé à un enracineur éliminant la plupart des dérivations morphologiques (e.g., « reliability »→« reliable »). Nous avons inclus un traitement morphologique simple supprimant la marque du pluriel soit le « s » final pour l'anglais (Harman, 1991).

Pour encore améliorer la qualité des réponses, nous avons également inclus une fonction de proximité entre termes (pour les détails, voir (Rasolofo et Savoy, 2003)). L'idée de base est la suivante. Si tous les termes de la requête apparaissent dans un document, le score attribué à cette page doit être augmenté. De plus, si ces termes se présentent proches les uns des autres ou s'ils apparaissent fréquemment, le score sera accru. À l'inverse, si seulement une partie de ces termes apparaissent, le score n'est pas modifié. Sachant que les requêtes soumises par les internautes sont très courtes (en moyenne 2,4 mots), ce calcul de proximité se limite souvent à contrôler si deux termes apparaissent bien dans un document donné.

Afin de savoir si l'emploi de cette composante s'avère bénéfique ou non, nous avons inclus le paramètre δ dans notre modèle. Si $\delta = 0$, cette fonction de proximité est ignorée ; par contre, si $\delta > 0$, la valeur de la fonction de proximité entre termes, multipliée par δ, sera ajoutée au score de la page Web calculé selon l'équation 1.

3. Dépistage des pages d'accueil et services en-ligne

Le dépistage d'une page unique correspond à une demande fréquente sur Internet. Ainsi, la réponse appropriée aux requêtes « sncf », « postes ouverts à la Cour suprême », « barbara mikulski », ou « observatoire du Mont-Blanc » ne correspond pas un ensemble de pages Web mais à accès direct au service (réservation en-ligne, consultation d'horaires), à la liste (films à l'affiche, postes ouverts) ou à la page d'accueil ou personnelle spécifiée.

3.1. *Notre modèle de recherche*

Afin d'atteindre une grande précision, notre moteur de recherche s'appuie sur le modèle Okapi avec trois représentations pour chaque document (voir section 2). La première indexation s'appuie sur l'ensemble du texte d'une page Web, ensemble comprenant également les phrases comprises dans les balises <TITLE> et les étiquettes <META>.

Cette première indexation repose uniquement sur le contenu d'une page et elle ignore les hyperliens indiquant des relations sémantiques entre pages. Par exemple, différentes pages à travers le monde pointent vers le département d'informatique de l'Université de Montréal (www.iro.umontreal.ca) et ces hyperliens sont indiqués dans ces pages sous la forme « la <href = "www.iro.umontreal.ca"> linguistique computationnelle possède un degré … ». Nous avons fait l'hypothèse que le texte encadré par la balise d'ancrage de l'hyperlien (soit « linguistique computationnelle » dans l'exemple précédent) décrit de manière concise et précise le contenu sémantique de la page pointée. Notons que ce texte possède aussi l'avantage d'être écrit par d'autres personnes que l'auteur de la page référencié, autorisant ainsi un enrichissement du vocabulaire d'indexation. En s'appuyant sur cette hypothèse, nous avons construit une deuxième indexation de chaque document sur la base de tous les textes d'ancrages d'une part et, d'autre part, du champ <TITLE> des pages pointant vers la page courante. Ainsi si la page P_k pointe vers P_i, le titre de P_k ainsi que le texte d'ancrage annonçant la présence d'un lien vers P_i sera inclus dans la deuxième représentation de P_i. Comme la page P_k n'est pas unique, nous concaténons les éléments extraits des pages P_k pour construire le représentant de P_i. Notre hypothèse s'appuie des études précédentes démontrant que les textes d'ancrage des hyperliens fournissent habituellement une bonne description de la page pointée (Craswell *et al.*, 2001 ; Westerveld *et al.*, 2002 ; Kraaij *et al.*, 2002).

Comme troisième indexation, nous avons concaténé le texte encadré par les balises <TITLE>, <H1> et <BIG> de toutes les pages P_k pointant vers la page courante P_i. Cette indexation renforce le rôle tenu par les pages faisant référence au document indexé.

3.2. *Évaluation*

Afin d'évaluer notre moteur de dépistage des pages d'accueil et de services en-ligne, nous avons utilisé la collection test de TREC 2003 comprenant des sites extraits du domaine « .gov » (1'247'753 pages pour un volume de 18,1 GB). En plus de ces pages, 300 requêtes avec leur réponse exacte sont incluses dans cette collection. Pour être précis, nous avons 352 bonnes réponses pour les 300 requêtes, soit en moyenne 1,173 bonne réponse par requête (médiane : 1, maximum : 6).

Comme mesure de performance, nous avons retenu l'inverse du rang moyen de la première réponse exacte retournée par la machine (mesure notée IRM). Ainsi, si le système place la bonne réponse en deuxième position, il obtient 1/2 point. Par contre, si cette réponse exacte apparaissait en 5e rang, 1/5 de point sera accordé.

La table 1 résume l'ensemble des expériences que nous avons menées. Dans la première colonne, nous avons regroupé les paramètres indiquant l'importance relative des trois représentations (α, β, γ) ainsi que le paramètre pondérant le calcul de proximité entre termes (δ). La deuxième colonne indique la valeur de l'IRM pour les 300 requêtes de notre évaluation (maximum 1 ou 100 %, minimum 0). La dernière indique le nombre de bonnes réponses dans les dix premiers rangs (sur un maximum 300) ainsi que le pourcentage correspondant.

Paramètres	IRM	dans les 10
$\alpha = 1$; $\beta = 0$; $\gamma = 0$; $\delta = 0$	0,335	163 (54,3%)
$\alpha = 0{,}6$; $\beta = 0{,}4$; $\gamma = 0$; $\delta = 0$	0,666	252 (84%)
$\alpha = 0{,}6$; $\beta = 0{,}4$; $\gamma = 0$; $\delta = 0{,}1$	0,691	251 (83,7%)
$\alpha = 0{,}6$; $\beta = 0{,}4$; $\gamma = 0{,}05$; $\delta = 0$	0,707	258 (86%)
$\alpha = 0{,}6$; $\beta = 0{,}4$; $\gamma = 0{,}05$; $\delta = 0{,}1$	**0,720**	**259** (86,3%)
$\alpha = 0{,}7$; $\beta = 0{,}3$; $\gamma = 0{,}05$; $\delta = 0{,}1$	0,700	258 (86%)
$\alpha = 0{,}8$; $\beta = 0{,}2$; $\gamma = 0{,}05$; $\delta = 0{,}1$	0,676	252 (84%)

Tableau 1. Évaluation de différentes variantes de notre stratégie de recherche (300 requêtes)

Si l'on ignore les hyperliens, nous devons indexer chaque page en fonction uniquement de son contenu. La performance d'une telle stratégie s'avère décevante (voir la deuxième ligne de la table 1 soit « $\alpha = 1$; $\beta = 0$; $\gamma = 0$; $\delta = 0$ »). La bonne réponse est retrouvée pour seulement 163 requêtes sur 300 (soit 54,3 %) si l'usager consulte les dix premières réponses extraites. La meilleure stratégie de recherche consiste à accorder une importance relativement marquée au contenu de la page elle-même (paramètre $\alpha = 0{,}6$) ainsi qu'à l'indexation basée sur les textes d'ancrage et champ <TITLE> des pages pointant vers la page considérée (paramètre $\beta = 0{,}4$). La présence de notre troisième représentation n'augmente que marginalement l'efficacité et la valeur de son paramètre ($\gamma = 0{,}05$) reste faible par rapport aux deux autres paramètres. L'usage de notre fonction de proximité ($\delta = 0{,}1$) ne permet pas d'augmenter le nombre de bonnes réponses trouvées (dernière ligne), mais la valeur IRM (deuxième colonne) augmente lorsque $\delta = 0{,}1$, signalant que l'ordinateur classe mieux les bonnes réponses qu'il a trouvées.

4. Extraction de ressources clés

En réponse à des requêtes telles que « virus informatique », « avantages et inconvénients de la légalisation de la marijuana ? » ou « Corée du Nord », les usagers désirent parfois obtenir une liste de ressources clés sur le thème spécifié. Il s'avère assez difficile de donner une définition opérationnelle d'une telle ressource. Certes, une page dont le contenu correspond au thème spécifié représente certainement une ressource à retenir et que l'on peut extraire à l'aide d'un

moteur de recherche classique. De bonnes pages centrales (c'est-à-dire des pages pointant vers des pages dont le contenu correspond au thème recherché) constituent aussi des ressources clés. Ainsi, au lieu de retourner plusieurs pages possédant la même page parente, il s'avère opportun d'extraire uniquement cette page parente. De manière similaire, si un site propose plusieurs pages possédant un contenu approprié, il est plus judicieux de retourner la page d'accueil du site ou celle correspondante à l'un de ses sous-sites.

4.1. *Notre modèle d'extraction de ressources clés*

Le modèle de recherche que nous proposons repose aussi sur trois indexations indépendantes des pages Web. Le premier représentant est identique au modèle précédent et regroupe tout le texte de la page à indexer (avec les balises <TITLE> et <META>). Cette stratégie ignore complètement les hyperliens. Une bonne ressource peut aussi être identifiée par les pages qui pointent vers elles d'une part, et d'autre part, par les textes d'ancrage des hyperliens quittant la page courante. Par exemple, si la page P_k pointe vers P_i qui pointe vers P_j, lors de l'indexation de P_i nous incluons les textes d'ancrage de P_k vers P_i de même que les textes d'ancrage dans P_i allant vers P_j. Enfin, le troisième représentant correspond à la concaténation de tous les champs <TITLE> et <H1> extraits des pages pointées par la page courante (c'est-à-dire de toutes les pages P_j de notre exemple précédent).

Afin d'évaluer la performance d'un système capable d'extraire les ressources clés sur un thème donné, nous avons retenu la précision obtenue après l'extraction de cinq éléments (notée « Précision à 5 ») et celle atteinte après la présentation de dix pages (indiquée sous la colonne « Précision à 10 »). Ces mesures indiquent le pourcentage de ressources clés extraites après cinq ou dix éléments retournés. Afin de mesurer ces précisions à l'aide de notre corpus « .gov », nous avons 50 requêtes avec leurs 516 ressources clés (soit, en moyenne, 10,32 ressources par requête, médiane : 8, maximum : 86, minimum : 1).

Paramètres	Précision à 5	Précision à 10
$\alpha = 1;\ \beta = 0;\ \gamma=0;\ \delta = 0$	8,00	6,20
$\alpha = 0{,}5;\ \beta = 0{,}5;\ \gamma = 0;\ \delta = 0$	16,40	10,80
$\alpha = 0{,}5;\ \beta = 0{,}5;\ \gamma = 0;\ \delta = 0{,}1$	16,00	11,00
$\alpha = 0{,}5;\ \beta = 0{,}5;\ \gamma = 0{,}03;\ \delta = 0$	16,40	10,80
$\alpha = 0{,}5;\ \beta = 0{,}5;\ \gamma = 0{,}03;\ \delta = 0{,}1$	16,00	11,00
$\alpha = 0{,}5;\ \beta = 0{,}5;\ \gamma = 0{,}1;\ \delta = 0$	15,60	11,40
$\alpha = 0{,}5;\ \beta = 0{,}5;\ \gamma = 0{,}1;\ \delta = 0{,}1$	16,00	**11,60**
$\alpha = 0{,}7;\ \beta = 0{,}3;\ \gamma = 0;\ \delta = 0$	15,20	10,20
$\alpha = 0{,}7;\ \beta = 0{,}3;\ \gamma = 0;\ \delta = 0{,}1$	16,00	10,20
$\alpha = 0{,}7;\ \beta = 0{,}3;\ \gamma = 0{,}1;\ \delta = 0$	14,00	11,40
$\alpha = 0{,}7;\ \beta = 0{,}3;\ \gamma = 0{,}1;\ \delta = 0{,}1$	14,00	11,40

Tableau 2. Évaluation de différentes stratégies d'extraction (50 requêtes, TREC 2003)

Si nous disposons uniquement du contenu d'une page Web afin de l'indexer, la précision à 5 s'élève à 8 % (ou à 6,2 % pour la précision à 10) (voir deuxième ligne de la table 2, « $\alpha = 1$; $\beta = 0$; $\gamma=0$; $\delta = 0$ »). Comme dans notre modèle précédent, il convient d'utiliser deux ($\alpha > 0$ et $\beta > 0$) ou trois ($\alpha > 0$, $\beta > 0$ et $\gamma > 0$) représentations afin d'améliorer la performance comme l'indique la table 2. Dans les grandes lignes, le meilleur choix se situe aux environs des valeurs « $\alpha = 0{,}5$; $\beta = 0{,}5$; $\gamma = 0{,}1$; $\delta = 0{,}1$ ». Le recours à la fonction de proximité (δ) n'apporte qu'une modification marginale de la performance. En effet, dans le cas présent, il ne s'agit plus de dépister une réponse unique mais plusieurs et cette fonction ne possède pas à elle seule la capacité de distinguer les bonnes réponses. Le troisième représentant (champ

<TITLE> et <H1> des pages pointées) n'apporte souvent qu'une légère amélioration de la précision à 10.

4.2. *Activation propagée, HITS et PageRank*

Pouvons-nous améliorer cette précision qui demeure relativement faible ? Comme une ressource clé est également un bon point de départ à la navigation, nous avons utilisé les listes de résultats obtenues précédemment afin de tenir compte des hyperliens d'une deuxième manière. Comme première stratégie, nous pouvons recourir à l'activation propagée (Savoy et Picard, 2001) qui transmet une fraction du score (notée λ) de chacune des pages extraites du Web vers ses voisines. Après avoir donné la requête à un moteur de recherche (dans notre cas, selon le modèle décrit en section 4.1), nous obtenons une liste triée de pages Web. Sur la base de cette liste, nous propageons les scores des documents selon les hyperliens sans tenir compte de leur orientation. Cette étape achevée, nous pouvons calculer le nouveau score pour chaque document en fonction de l'équation suivante :

$$RSV'(D_i) = RSV(D_i) + \lambda \cdot \sum_{j=1}^{r} RSV(D_j) \qquad (2)$$

dans laquelle $RSV'(D_i)$ indique le nouveau score de la page D_i dépendant de son score initial (noté $RSV(D_i)$, calculé par l'équation 1) et du score de ses r voisins. Dans une version plus sophistiquée de cette stratégie, le facteur de propagation λ pourrait dépendre du type de lien reliant les deux documents ou du contenu respectif de D_i et de D_j. De plus, au lieu de propager une fraction du score vers tous les voisins, on peut limiter cette propagation au r « meilleurs » voisins, c'est-à-dire aux voisins les mieux classés par le moteur de recherche.

Comme deuxième stratégie d'extraction s'appuyant sur les hyperliens, nous pouvons recourir à l'algorithme HITS proposé par Kleinberg (1999). Dans ce modèle, on estime qu'une page pointant vers d'autres sources d'information est perçue comme une bonne « page centrale » (*hub*) tandis qu'un document possédant beaucoup d'hyperliens entrants sera analysé comme une bonne « page de référence » (*authority*). De manière récursive, si une page pointe vers plusieurs bonnes « pages de référence », elle obtiendra une meilleure évaluation en tant que « page centrale ». De manière duale, si une page est pointée par de nombreuses bonnes « pages centrales », sa valeur, comme « page de référence », augmentera.

Dans cette approche, chaque page D_i possède deux scores, soit celui comme « page de référence », noté $A^{c+1}(D_i)$, et celui comme « page centrale », noté $H^{c+1}(D_i)$, valeurs obtenues après c+1 itérations. Ces scores sont calculés selon les formules suivantes :

$$A^{c+1}(D_i) = \sum_{D_k \in parent(D_i)} H^c(D_k) \quad \text{et} \quad H^{c+1}(D_i) = \sum_{D_j \in fils(D_i)} A^c(D_j) \qquad (3)$$

dans lesquelles $parent(D_i)$ indique l'ensemble des pages pointant vers D_i et $fils(D_i)$ l'ensemble des pages pointées par D_i.

Comme pour l'activation propagée, cet algorithme débute par l'établissement d'un ensemble de pages jugées pertinentes par un moteur de recherche. De cette liste, on extrait les r meilleures pages (ensemble noté σ) ainsi que toutes les pages pointées (tous les fils de σ) et toutes les pages pointant vers l'un des éléments de σ (tous les ancêtres directs). Sur cet ensemble élargi, on calcule les scores selon l'équation 3 et on les normalise (division par la somme des carrées de toutes les valeurs). Lors de la phase initiale, on attribue une valeur unitaire aux deux scores

de chaque page. Finalement, on itère durant cinq cycles car après ces cinq étapes, le classement des documents ne se modifie généralement plus.

Brin et Page (1998) ont proposé un modèle nommé PageRank débutant également par une liste triée de pages. Cependant, au lieu de classer ces pages en fonction de leur score (selon l'équation 1), PageRank va réordonner cet ensemble de pages dépistées en fonction de leur connectivité. En présence de la page D_i ayant les documents D_1, D_2, … D_m pointant vers elle, nous calculons son score PageRank, noté $PR^{c+1}(D_i)$ au cycle c+1, par l'équation suivante :

$$PR^{c+1}(D_i) = (1\text{-}d) \cdot [1 / n] + d \cdot [(PR^c(D_1) / C(D_1)) + \ \dots \ + (PR^c(D_m) / C(D_m))] \qquad (4)$$

dans laquelle d est un paramètre (fixé à 0,85), n indique le nombre de documents dans la collection et $C(D_k)$ indique le nombre d'hyperliens sortant du document D_k.

L'idée sous-jacente est la suivante. La valeur PageRank d'une page correspond à une estimation de la probabilité qu'un internaute la rencontre en naviguant de manière aléatoire. Ce dernier peut, avec une probabilité (1-d), choisir n'importe quel document sur le Web (et chaque page possède la même probabilité 1/n d'être choisie). D'un autre côté, l'internaute peut, avec une probabilité d, choisir de suivre un des hyperliens qu'il voit sur la page courante et chaque hyperlien possède la même chance d'être sélectionné. Dans ce cas, la probabilité d'aboutir sur la page D_i dépendra donc de la probabilité d'être sur une des pages D_j (notée $PR^c(D_j)$) et de l'inverse du nombre d'hyperliens inclus dans cette page D_j (soit $C(D_j)$).

Comme dans l'algorithme de Kleinberg, le score PageRank des pages se calcule de manière itérative et, habituellement, on effectue cinq itérations. Après chaque itération, on normalise les valeurs $PR^c(D_i)$ en les divisant par la somme des valeurs $PR^c(D_j)$ et, comme valeurs initiales, on fixe $PR^0(D_i)$ à 1/n.

Paramètres	Précision à 5	Précision à 10
$\alpha = 0{,}5$; $\beta = 0{,}5$; $\gamma = 0{,}1$; $\delta = 0{,}1$	16,00	**11,60**
$\lambda = 0{,}02$; r = 50	17,60	12,00
$\lambda = 0{,}05$; r = 50	**17,60**	12,20
$\lambda = 0{,}1$; r = 50	13,60	12,00
$\lambda = 0{,}05$; r = 200	19,20	12,40
$\lambda = 0{,}05$; r = 300	16,40	**14,00**
$\lambda = 0{,}05$; r = 400	16,00	13,80
HITS, *hub*, $\sigma = 50$ premiers	3,60	2,60
HITS, *authority*, $\sigma = 50$ premiers	0.80	0.60
PageRank, d = 0,85	2,00	1,60

Tableau 3. Évaluation de stratégies utilisant les hyperliens (50 requêtes, TREC 2003)

Afin de mesurer la performance de ces trois solutions, nous avons repris, dans la première ligne de la table 3, la meilleure solution obtenue par notre modèle d'extraction (voir tableau 2). Sur cet ensemble ordonné de pages, nous avons appliqué notre algorithme d'activation propagée en faisant varier le facteur de propagation λ ainsi que le nombre r de pages sur lesquelles cette propagation est appliquée. De même, nous avons extrait les 50 premières pages pour former l'ensemble σ , puis nous avons élargi cet ensemble et avons calculé le score des pages centrales (*hub*) et celui des pages de référence (*authority*). Enfin, nous avons utilisé l'algorithme PageRank. Les résultats de la table 3 sont clairs. Les algorithmes HITS ou PageRank n'apportent pas la réponse souhaitée.

5. Conclusion

Différentes études ont démontré que les hyperliens ne permettaient pas, ou de façon marginale, d'améliorer la qualité des moteurs de recherche sur le Web (Savoy et Picard, 2001). Or, si le type d'interrogation correspond à la recherche de service en-ligne (ou de page d'accueil) d'une part, ou, d'autre part, à l'extraction de ressources clés, l'inclusion des textes d'ancrage dans l'indexation des pages Web permet d'augmenter significativement les performances obtenues, du moins dans le cadre de notre proposition d'indexation combinée. Dans le cadre du dépistage de ressources clés, l'algorithme HITS ou PageRank n'apporte pas une solution intéressante tandis que notre approche basée sur le principe de l'activation propagée permet une augmentation de la précision de l'ordre de 10 à 20 %.

Références

Brin S. et Page L. (1999). The anatomy of a large-scale hypertextual Web search engine. In Mendelzon A.(Ed.), *Proceedings of the WWW8* : 107-117.

Craswell N., Hawking D. et Robertson S. (2001). Effective site finding using link anchor information. In Kraft D.H., Croft W.B., Harper D.J. et Zobel J. (Eds), *Proceedings of ACM-SIGIR'2001* : 250-257.

Harman D. (1991). How effective is suffixing? *Journal of the American Society for Information Science,* vol. (42/1) : 7-15.

Kleinberg J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, vol. (46/5) : 604-632.

Kraaij W., Wes/terveld T. et Hiemstra D. (2002). The importance of prior probabilities for entry page search. In Baeza-Yates R., Beaulieu M., Järvelin K. et Myaeng S.H. (Eds), *Proceedings of ACM-SIGIR'2002* : 27-34.

Nielsen J. (2000). *Designing Web usability: the practice of simplicity*. New Riders.

Radev D.R., Libner K. et Fan W. (2002). Getting answers to natural language questions on the Web. *Journal of the American Society for Information Science and Technology*, vol. (53/5) : 359-364.

Rasolofo Y. and Savoy J. (2003). Term proximity scoring for keyword-based retrieval systems. In Sebastiani F. (Ed.), *Proceedings of ECIR 2003* : 207-218.

Robertson S.E., Walker S. et Beaulieu M. (2000). Experimentation as a way of life: OKAPI at TREC. *Information Processing & Management*, vol. (36/1) : 95-108.

Savoy J. et Picard J. (2001). Retrieval effectiveness on the Web. *Information Processing & Management*, vol. (37/4) : 543-569.

Savoy J. et Rasolofo Y. (2003). Report on the TREC-11 experiment: Arabic, named page and topic distillation searches. In Voorhees E.M. et Buckland L.P. (Eds), *Proceedings of TREC 2002* : 765-774.

Spink A., Wolfram D., Jansen M.B.J. et Saracevic T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology*, vol. (52/3) : 226-234.

Watters C. et Amoudi G. (2003). GeoSearcher: location-based ranking of search results. *Journal of the American Society for Information Science and Technology*, vol. (54/2) : 140-151.

Westerveld T., Kraaij W. et Hiemstra D. (2002). Retrieving Web pages using content, links, URLs and anchors. In Voorhees E.M. et Harman D.K. (Eds), *Proceedings of TREC 2001* : 663-672.

Hypothèses pour la construction et l'exploitation conjointe d'une base lexicale sémantique basée sur les vecteurs conceptuels

Didier Schwab, Mathieu Lafourcade, Violaine Prince

LIRMM – Laboratoire d'informatique, de Robotique
et de Microélectronique de Montpellier
Montpellier – France
{schwab, lafourcade, prince}@lirmm.fr

Abstract

In the framework of meaning representation in NLP, the TAL research group at LIRMM develops a system for analysing thematic aspects of texts and for lexical disambiguation based on conceptual vectors. These vectors aim at representing sets of ideas related to any kind of textual segment (words, phrases, sentences, texts). Associated to various information, lexical functions for instance, vectors allow to represent the semantic of such textual segments. In this paper, we expose the hypothesis grounding the construction of a semantic lexical database. We present the overall organisation of such a system and an example of cooperation between agents in the context of a lexical item learning.

Résumé

Dans le cadre de la représentation du sens en TALN, l'équipe TAL (Traitement Algorithmique des Langues) du LIRMM développe un système d'analyse des aspects thématiques des textes et de désambiguïsation lexicale basée sur les vecteurs conceptuels. Ces vecteurs cherchent à représenter un ensemble d'idées associées à tout segment textuel (mots, expressions, textes,...). Associés à des informations de nature diverse, par exemple des fonctions lexicales, les vecteurs peuvent permettre de représenter la sémantique de ces segments. Dans cet article, nous exposons les hypothèses de départ que nous avons prises pour la construction d'une base lexicale sémantique. Nous présentons l'implémentation réalisée sous forme de système multi-agents et accessible en ligne ainsi que quelques expérimentations existantes ou en cours de réalisation.

Keywords: conceptual vectors, sense representation, semantic lexical base, construction hypothesis, lexical functions, multi-agents systems.

Mots-clés : vecteurs conceptuels, représentation du sens, base lexicale sémantique, hypothèses de constructions, fonctions lexicales, systèmes multi-agents.

1. Introduction

Dans le cadre de la représentation du sens en Traitement Automatique du Langage Naturel (TALN), l'équipe Traitement Algorithmique des Langues (TAL) du LIRMM développe actuellement un système d'analyse des aspects thématiques des textes et de désambiguïsation lexicale basée sur les vecteurs conceptuels. Ces vecteurs cherchent à représenter un ensemble d'idées associées à tout segment textuel (mots, expressions, textes,...). Associés à des informations lexicales, les relations sémantiques par exemple, les vecteurs peuvent permettre de représenter la sémantique de ces segments. Dans cet article, nous exposons les hypothèses de départ que nous avons prises pour la construction d'une base lexicale sémantique, c'est-à-dire les moyens utilisés pour fabriquer des objets appelés *acception* permettant de représenter le sens d'un maxi-

mum de termes de la langue. Ces hypothèses, au nombre de cinq, sont (I) *représentation du sens par une approche combinant approche thématique (vectorielle) et approche lexicale (relations sémantiques externes)*, (II) prise en compte des *relations sémantiques internes* (polysémie), (III) *génération automatique* des acceptions, (IV) réalisation d'une *analyse multi-sources* (à partir de dictionnaires, listes de synonymes, sites web,...) et (V) *apprentissage permanent*. Ces hypothèses nous ont conduit à adopter une architecture multi-agents dont nous présentons ensuite succinctement l'implémentation accessible en ligne. Nous conclurons cet article par un exemple de collaboration entre agents ainsi que sur quelques expérimentations existantes ou en cours de réalisation.

2. Objectifs : approches polyvalentes

L'utilisation des vecteurs conceptuels se fait dans le cadre d'un système de TALN qui a pour objectif la représentation du sens et la création d'outils exploitant cette représentation. Les applications de ce système visent des domaines aussi variés que la recherche d'informations, la traduction ou le résumé automatique. Dès lors, nous développons un système qui doit être le plus générique et le plus évolutif possible. Il tente de représenter un maximum de termes du lexique des langues traitées, ce qui pose un problème d'acquisition des connaissances : un problème d'acquisition de ce lexique (acquérir un maximum d'*items lexicaux*[1] constituant ce lexique) mais aussi un problème d'acquisition du sens de ces items (en récupérant des données qui concernent le sens et en les exploitant). C'est cette double problématique que nous tentons de résoudre dans le cadre de ce projet.

3. Hypothèses de construction d'une base sémantique lexicale

L'acquisition du sens des termes est basée sur cinq hypothèses fondamentales. C'est sur elles que reposent les architectures conceptuelles et implémentationnelles que nous avons choisies pour notre système.

3.1. *Hypothèse I. Représentation du sens : approche combinant représentation thématique et informations lexicales*

3.1.1. *Vecteurs conceptuels*

Nous représentons les aspects thématiques des segments textuels (documents, paragraphes, syntagmes, etc.) par des vecteurs conceptuels. Si les vecteurs ont été utilisés en informatique documentaire pour la recherche d'information dès la fin des années 1960 (Salton, 1968), leur emploi pour la représentation du sens est plus le fait du modèle LSI (*Latent Semantic Indexing* (Deerwester *et al.*, 1990)) issu de l'analyse sémantique latente en psycholinguistique. En informatique, et de façon presque concurrente, c'est à partir de Chauché (1990) que l'on a une formalisation de la projection de la notion, linguistique cette fois, de champ sémantique dans un espace vectoriel. À partir d'un ensemble de notions élémentaires données *a priori* dont nous faisons l'hypothèse, les concepts (dans notre expérimentation sur le français nous utilisons le *Larousse* (1992) dans lequel sont définis 873 concepts), il est possible de construire des vecteurs (dits conceptuels) et de les associer à des items lexicaux. Les termes polysémiques combinent les différents vecteurs correspondant aux différents sens. Cette approche vectorielle est fondée sur des propriétés mathématiques bien connues sur lesquelles il est possible d'effectuer des ma-

[1] Les items lexicaux sont des mots ou des expressions qui constituent les entrées du lexique. Par exemple, ‹*voiture*› ou ‹*pomme de terre*› sont des items lexicaux. Dans la suite de cet article, par abus de langage, nous utiliserons parfois mot ou terme pour qualifier un item lexical. Nous noterons les items en minuscule et entre apostrophes (‹*vie*›) et les concepts en majuscules (VIE).

nipulations formellement pertinentes auxquelles sont rattachées des interprétations linguistiques appropriées. L'hypothèse principale du thésaurus, que nous adoptons ici, est que cet ensemble constitue un espace générateur pour les termes et leurs sens. De manière plus générale, n'importe quel sens peut s'y projeter selon le principe suivant :

Soit $\mathcal{C}$ un ensemble fini de n concepts. Un vecteur conceptuel V est une combinaison linéaire d'éléments c_i de $\mathcal{C}$. Pour une idée A, le vecteur V_A est la description en extension des activations de tous les concepts de $\mathcal{C}$. Par exemple, les différents sens d'‹*existence*› peuvent être projetés sur les concepts suivants (les CONCEPT[intensité] sont ordonnés par valeurs décroissantes) : V‹*existence*› = (EXISTENCE[0.82], VIE[0.44], IDENTITÉ[0.38], ÉTAT[0.33],...). En pratique, plus $\mathcal{C}$ est grand, plus fines sont les descriptions de sens mais plus leur manipulation est lourde. Il est clair que pour les vecteurs denses, ceux qui ont peu de coordonnées nulles, l'énumération des concepts activés est longue et la pertinence difficile à évaluer. En général, pour évaluer la qualité d'un vecteur, nous préférerons sélectionner les termes thématiquement proches, le *voisinage* (noté $\mathcal{V}$). Par exemple, pour ‹*vie*› : $\mathcal{V}$(‹*existence*›) : ‹*existence*›, ‹*exister*›, ‹*vivant*›, ‹*vie*›,... Cette opération est réalisée à l'aide de la distance angulaire.

Soit $Sim(X, Y)$ une des mesures de *similarité* entre deux vecteurs X et Y, souvent utilisée en recherche d'information (Morin, 1999). $Sim(X, Y) = \cos(\widehat{X, Y}) = \frac{X \cdot Y}{\|X\| \times \|Y\|}$ avec "·" désignant le produit scalaire. Nous supposons ici que les composantes des vecteurs sont positives ou nulles, la *distance angulaire* entre deux vecteurs X et Y est $D_A(X, Y) = \arccos(Sim(X, Y))$. Intuitivement, cette fonction constitue une évaluation de la *proximité thématique* et en pratique la mesure de l'angle entre les deux vecteurs. Nous considérons en général que pour une distance $D_A(X, Y) \leq \frac{\pi}{4}$ (45°), X et Y sont thématiquement proches et partagent plusieurs concepts. Pour $D_A(X, Y) > \frac{\pi}{4}$, la proximité thématique est considérée comme faible et aux alentours de $\frac{\pi}{2}$ (90°), X et Y n'ont aucune relation. On remarquera que ces seuils ne servent que d'indicateurs pour un réviseur humain et restent à la fois subjectifs et arbitraires. D_A est une distance, elle vérifie donc les propriétés de réflexivité, de symétrie et d'inégalité triangulaire. Nous obtenons, par exemple, les angles suivants [2] :

D_A(V(‹*locomotive*›),V(‹*locomotive*›))=0 (0°)
D_A(V(‹*locomotive*›),V(‹*locomotrice*›))=0.24 (14°)
D_A(V(‹*locomotive*›),V(‹*automotrice*›))=0.22 (13°)
D_A(V(‹*locomotive*›),V(‹*rhododendron*›))=1.15 (65°)
D_A(V(‹*locomotive*›),V(‹*train*›))=0.54 (31°)
D_A(V(‹*locomotive*›),V(‹*guépard*›))=0,94 (54°)

Le premier résultat a une interprétation directe, ‹*locomotive*› ne peut être plus proche d'autre chose que de lui même. Les termes ‹*automotrice*› et ‹*locomotrice*› sont quasi-synonymes de ‹*locomotive*›, ce qui explique les deux résultats suivants. Le peu de rapport entre ‹*locomotive*› et ‹*rhododendron*› explique l'écart entre leur vecteur respectif. Dans le dernier exemple, l'angle peu important entre ‹*locomotive*› et ‹*guépard*› au regard de celui entre ‹*locomotive*› et ‹*rhododendron*› se comprend si l'on se rappelle que D_A est une distance thématique et non une distance ontologique ou sémantique. Les deux items ont en commun de partager une idée de rapidité. On remarquera que les comparaisons entre les valeurs sont plus significatives que les valeurs elles-mêmes.

3.1.2. *Informations lexicales : fonctions lexicales*

Nous venons de le voir, les vecteurs conceptuels constituent une représentation thématique et non une représentation sémantique. En psycholinguistique, il est aujourd'hui bien accepté que le sens est issus à la fois du niveau des vocables et du niveau des thèmes (Kawamoto, 1993). Des informations spécifiques de rapports de sens que les termes entretiennent entre eux sont

[2] Les exemples sont extraits de http//www.lirmm.fr/~schwab

donc indispensables à notre modèle. Ces rapports peuvent être modélisés sous la formes de fonctions lexicales à la Mel'čuk. Par exemple, ‹*autoriser*› entretient une relation d'antonymie avec ‹*interdire*› ou ‹*défendre*› et une relation de synonymie avec ‹*permettre*›.

3.2. *Hypothèse II. Relations sémantiques internes d'un item lexical*

Les *relations sémantiques internes* sont les diverses relations qu'il existe entre les différents sens d'un même item lexical. Ces relations concernent ce que l'on appelle *l'aire sémantique* d'un mot, c'est-à-dire l'ensemble des significations qu'il est susceptible de prendre. On distingue la *monosémie* qui concerne les termes qui n'ont qu'un seul sens (‹*calame*›, ‹*cajou*›, ‹*neuroleptique*›,. . .), et la *polysémie* qui concerne les termes qui ont plusieurs sens[3] (‹*louer*›, ‹*froid*›, ‹*frégate*›,. . .).

Les relations sémantiques internes d'un item lexical, le fait qu'il puisse avoir plusieurs sens nous obligent à en tenir compte dans la construction de nos vecteurs conceptuels. Nous pensons qu'il est difficile de considérer qu'une quelconque tâche de désambiguïsation est possible sans que chaque sens d'un item lexical ne soit aussi affecté d'un vecteur conceptuel (chacun de ces vecteurs entrant dans la construction du vecteur global sans doute de façon non-linéaire).

Nous appelons *acception* un sens particulier d'un item lexical admis et reconnu par l'usage. Il s'agit d'une unité sémantique propre à une langue donnée (Sérasset et Mangeot, 2001). Par exemple, l'item lexical ‹*botte*› a au moins trois acceptions : la *« chaussure »*, l'*« amas de végétaux »* ou le *« coup en escrime »*. Contrairement aux items lexicaux, les acceptions sont donc monosémiques.

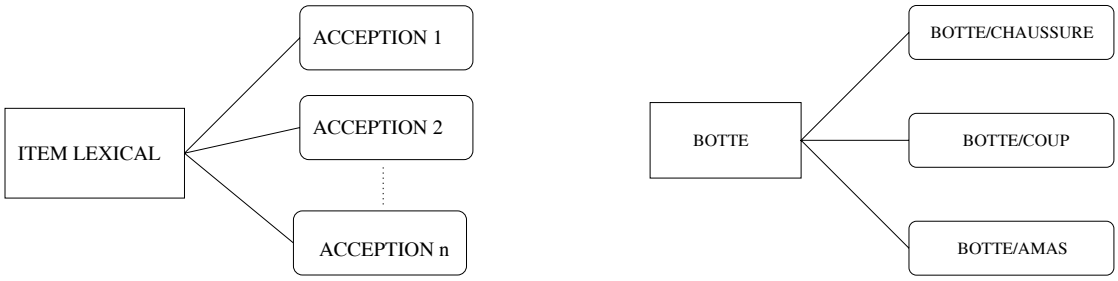
Nous avons tenu compte de ces deux premières hypothèses pour construire les objets acceptions. Les acceptions sont composées d'un certain nombre d'**informations linguistiques** comme la **morphologie** (composé des **catégories grammaticales** : *nom, pronom, adjectif, verbe*, etc., du **genre** : *masculin, féminin, neutre* et le *nombre* : *singulier, pluriel*), la **fréquence en usage** (le nombre de fois ou au moins une évaluation où l'acception a été rencontrée), **un vecteur conceptuel**, des **fonctions lexicales**, des **informations étymologiques**, des **gloses** (des informations que l'on trouve, par exemple, dans les dictionnaires de traduction ou de synonymie pour préciser le sens d'un mot).

3.3. *Hypothèse III. Génération automatique*

Notre objectif est donc de construire une base de stockage d'objets acception. La difficulté principale vient de la création de ces objets. Dans des dictionnaires francophones classiques comme le *Larousse* (2001) et le *Hachette* (2000), sur les 80000 items (mots communs, nom propres, expressions,. . .) répertoriés, une majorité est polysémique. Dans notre expérience sur le français, pour un peu plus de 100000 entrées (mots communs, nom propres, expressions,. . .), le taux de termes polysémiques est d'environ 60%. Le nombre moyen de définitions pour ces derniers étant d'un peu plus de 5, il faudrait indexer à la main plus de 400000 acceptions, ce qui paraît irréaliste.

Nous considérons qu'il est possible d'automatiser cette tâche grâce à un apprentissage basé sur des informations extraites de diverses sources (dictionnaires, listes de synonymes, indexations manuelles, recherches web,. . .). Pour chaque item lexical, chacune de ces sources peut nous

[3] Nous confondons volontairement ici l'homonymie et la polysémie. Le problème de leur différentiation est assez peu intéressant au niveau du TALN en général et de la désambiguïsation en particulier. Notre but est de trouver le meilleur sens pour un terme et non de savoir comment le terme a acquis ce nouveau sens ni quels sont les rapports qu'entretiennent ces sens entre eux. Nous parlerons dans la suite de cet article d'items polysémiques sans distinguer s'il s'agit d'une vraie polysémie ou d'un cas d'homonymie.



(a) Organisation générale de la représentation du sens pour un item lexical. Le vecteur conceptuel général d'un item est calculé à partir des vecteurs de chacune des acceptions de cet item.

(b) Organisation générale de la représentation du sens pour l'item lexical ‹*botte*›

Figure 1. Organisation générale de la représentation du sens pour un item lexical.

donner une ou plusieurs définitions. Il est possible d'extraire un certain nombre d'informations de ces définitions. Les plus riches sont celles extraites des dictionnaires, elles peuvent permettre d'obtenir des informations d'ordre morphologique et éventuellement étymologique. À partir du texte des définitions, il est aussi possible de calculer un vecteur conceptuel. L'objet qui regroupe toutes les informations qu'il est possible d'extraire d'une définition est appelé *lexie*. Sa structure interne est identique à celle d'une acception (cf. hypothèse I). L'essentiel de l'apprentissage principal se fait sur des dictionnaires à usage humain (dictionnaires classiques, de synonymes, d'antonymes,...) ou machinaux. Dans la perspective où nous nous plaçons, dans un espace dimensionné en fonction d'une hiérarchie de concepts, un amorçage du système d'apprentissage est nécessaire. Il s'agit d'affecter des vecteurs à un nombre réduit d'entités qui sont choisies en fonction de leur fréquence en langue et/ou de leur polysémie. La taille du noyau est très réduite (un millier de termes environ) et les éléments de ce noyau considérés comme pertinents. À partir de ce noyau, le processus d'apprentissage peut débuter. La méthode d'analyse construit, à partir de vecteurs conceptuels déjà existants et de nouvelles définitions, de nouveaux vecteurs. L'idée est qu'à partir d'un noyau réduit d'items pertinents, un apprentissage sur des définitions permet de créer une cohérence entre les vecteurs et donc de générer une base d'items pertinents.

3.4. *Hypothèse IV. Analyse multi-sources*

L'analyse des définitions pose un certain nombre de problèmes quant à la lecture du sens. C'est le cas du métalangage, c'est-à-dire le langage utilisé pour structurer le dictionnaire. Ce dernier est aisément utilisable lorsqu'il s'agit de récupérer les catégories grammaticales des items mais certaines constructions de définitions sont difficilement compréhensibles sans compétence métalinguistique. Il existe un certain nombre de phrases qu'il convient d'analyser convenablement(*partie de*, *se dit de*, etc.) ou d'ambiguïtés que même un humain ne peut lever. Comment savoir, par exemple, que *en parlant* est du métalangage dans une définition de l'item ‹*aboyer*› comme *« Crier, en parlant du chien »*? Pour pallier de tels manques définitoires, nous utilisons diverses sources lexicales. Il s'agit de tempérer statistiquement les diverses incohérences locales. Ainsi, si une définition est mal formée (donc difficilement analysable correctement), une autre définition, mieux formée et provenant d'une autre source, pourra corriger l'effet de la première.

Aucune source ne peut couvrir l'intégralité du lexique non seulement parce que celui-ci est en constance évolution mais surtout parce que rechercher, trouver et finalement décrire de façon systématique chaque terme de la langue est une tâche au mieux extrêmement difficile. Il fau-

drait, en effet, connaître l'ensemble des formes mais aussi l'ensemble des sens que peuvent prendre ses formes dans n'importe quel domaine d'activité, dans n'importe quelle structure sociale et quel que soit le niveau de langage. L'utilisation de multi-sources permet donc aussi de maximiser la probabilité de récupérer une définition pour certains termes qui pourraient être trouvés dans certains dictionnaires et pas dans d'autres. Par exemple, ‹*liturgiste*› ne se trouve pas dans le *Larousse* (2001) mais on le trouve dans le *Hachette* (2000).

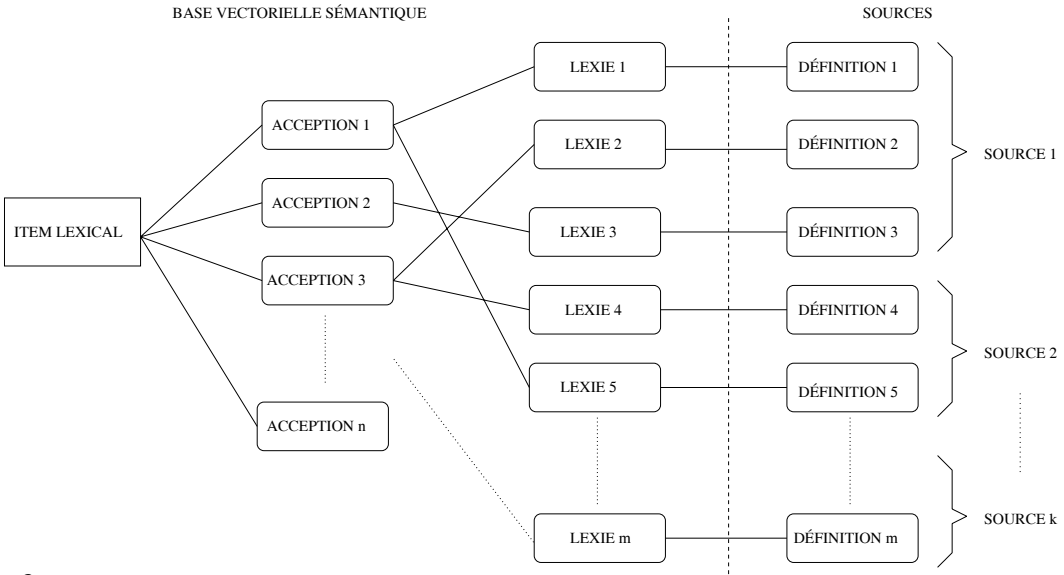


Figure 2. Ce schéma présente l'organisation générale de la représentation du sens pour un item lexical. À partir des informations données par les définitions, on crée des lexies. Ces lexies sont ensuite utilisées pour fabriquer les acceptions correspondant à chaque sens du mot.

La figure 2 présente la structuration générale de la base sémantique lexicale. À partir de diverses sources, on peut obtenir des définitions pour un item lexical. Pour chaque définition, on crée une lexie qui est un objet qui regroupe les diverses informations contenues dans cette définition dont un vecteur conceptuel calculé à partir du texte de la définition. Ces lexies doivent alors être catégorisées (regroupées par sens) afin de fabriquer les diverses acceptions de l'item.

Prenons un exemple. Considérons trois sources, le *Petit Larousse* (2001), le dictionnaire *Hachette de la langue* (2000) et l'*Encyclopédie Club-internet*[4]. Pour l'item ‹*botte*›, nous trouvons les définitions suivantes :

botte.1 : #nf# Réunion de végétaux de même nature liés ensemble. (Une botte de paille, de radis, de fleurs) . *[Hach]*

botte.2 : #nf# En escrime, coup porté à l'adversaire avec un fleuret ou une épée. (Pousser, porter, parer une botte) (Botte secrète.). *[Hach]*

botte.3 : #nf# Chaussure de cuir, de caoutchouc ou de plastique qui enferme le pied et la jambe, parfois la cuisse. (Des bottes de cavalier) – Chaussure d'extérieur basse. (Botte d'hiver, de ski, de marche) . *[Hach]*

botte.4 : #nf# (néerl. bote, touffe de lin) . Assemblage de végétaux de même nature liés ensemble : (Botte de paille. Botte de radis.) . *[Lar]*

botte.5 : #nf# (#ethym-it# botta, coup) . Coup de pointe donné avec le fleuret ou l'épée . *[Lar]*

botte.6 : #nf# (p.-ê. de bot) . Chaussure à tige montante qui enferme le pied et la jambe généralement jusqu'au genou : (Bottes de cuir) . *[Lar]*

botte.7 : #nf# (néerl. (bote) « touffe de lin »). Assemblage de végétaux, de même sorte, tenus par un lien. (Une botte de paille, d'asperges.). *[Club]*

[4] http//www.club-internet.fr

botte.8 : #nf# (ital. (botta) « coup »). En escrime, coup de pointe à l'épée ou au fleuret. (Porter une botte.) #fig# Attaque vive . *[Club]*

botte.9 : #nf# Chaussure montant jusqu'à mi-jambe ou jusqu'au genou, enfermant parfois une partie de la cuisse. (Des bottes de pêche, d'équitation.). *[Club]*

Il semble clair que les sens {1,4,7} peuvent se regrouper pour former le sens d'*« amas »*, les sens {2,5,8} pour former le sens de *« coup »* et les sens {3,6,9} pour former l'acception *« chaussure »*. La figure 3 présente l'organisation générale de l'item ‹*botte*› telle qu'elle serait si celui-ci était défini par ces neuf définitions.

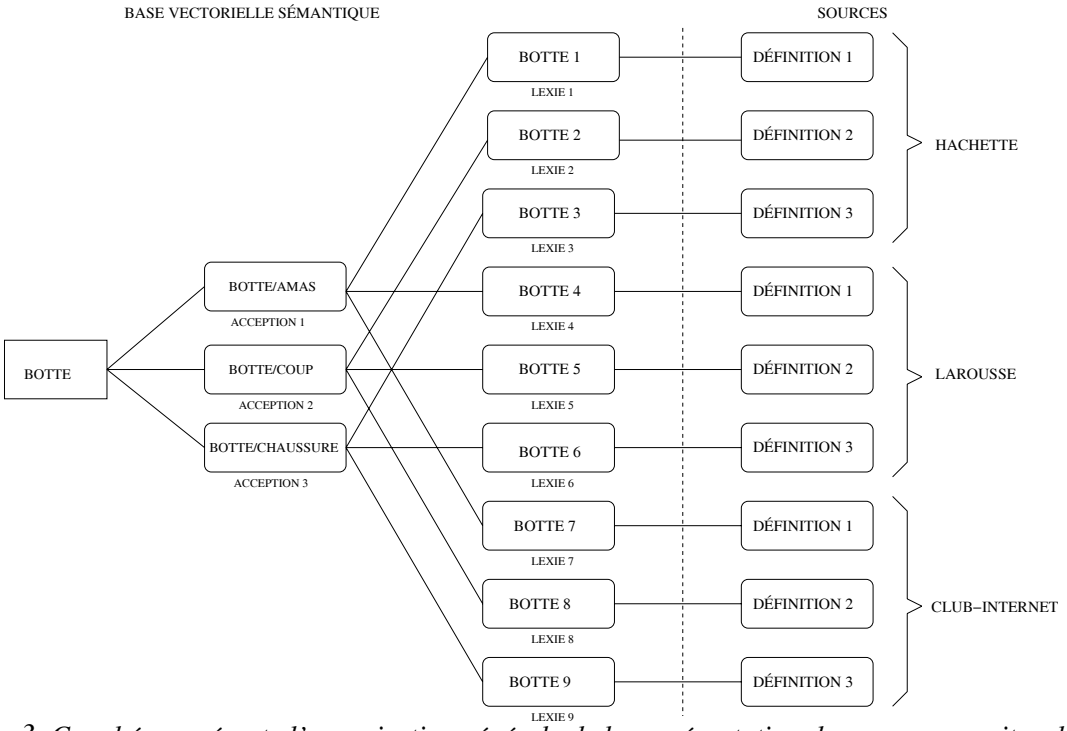


Figure 3. Ce schéma présente l'organisation générale de la représentation du sens pour un item lexical. À partir des informations données par les définitions, on crée des lexies. Ces lexies sont ensuite utilisées pour fabriquer les acceptions correspondant à chaque sens du mot.

3.5. *Hypothèse V. Apprentissage permanent*

Pour analyser un certain nombre de documents, en particulier les journaux, il est souvent nécessaire de savoir à quoi correspondent des néologismes, qui sont certaines personnalités ou encore quel est le domaine d'activité d'une entreprise. Par exemple, dans un texte, l'usage du nom de l'entreprise ‹*Arcelor*› indique vraisemblablement un contexte axé sur le traitement de l'acier. Les diverses sources et en particulier le web par les serveurs d'informations (*Le Monde, Libération,...*) peuvent permettre les informations nécessaires à la fabrications des structures appropriées.

De plus, il est difficile de penser (notre expérimentation nous l'a montré) que les vecteurs deviendraient cohérents dès la première passe. Il est vraisemblable que des mots clés d'une définition n'aient pas encore été appris par le système lors de son analyse. La convergence des vecteurs vers une position quasi-stable ne pourra se faire que dans un nombre de cycles qu'il est impossible de déterminer à l'avance mais qui est fonction de la taille et de la richesse du vocabulaire utilisé dans les définitions. Ces deux raisons, la variabilité lexicale et l'impossibilité

de véritablement stabiliser une base, nous ont conduits à considérer cette cinquième et dernière hypothèse : la base est en apprentissage permanent.

4. Vers une société d'agents

Nous venons de le voir, notre objectif est la création d'un véritable système permettant l'apprentissage d'informations sémantiques (thématiques par les Vecteurs Conceptuels et lexicales par les relations sémantiques) et leur exploitation. Il s'agit de récupérer des définitions, les analyser à l'aide des *lexies* et *acceptions* déjà calculés afin d'en fabriquer ou d'en réviser d'autres (cf. section 3). L'analyse simple des définitions peut ne pas toujours suffire à améliorer la cohérence des vecteurs (difficulté d'analyser certaines tournures, problèmes de métalangage,. . .). Plusieurs solutions sont alors possibles comme par exemple, l'utilisation des relations sémantiques existant entre les items (synonymie (Lafourcade et Prince, 2001), antonymie (Schwab *et al.*, 2002c), hypéronymie,. . .). Ces relations interviennent alors à deux niveaux : au niveau de la construction, pour fabriquer les acceptions et au niveau de l'exploitation puisqu'elles participent alors à une meilleure construction du sens.

L'exploitation de la base peut aussi être plurielle : utilisation pour la désambiguïsation sémantique, annotation, transfert lexical, traduction, recherche d'informations. À la fois pour l'exploitation et l'apprentissage, il est donc nécessaire de pouvoir facilement ajouter des modules apportant tel ou tel service. C'est une des raisons pour lesquelles notre vision de l'architecture nécessaire s'est rapidement rapprochée des systèmes multi-agents (SMA).

4.1. *SMA et TALN*

Les SMA sont issus de l'Intelligence Artificielle Distribuée (IAD) qui répartit l'intelligence dans des agents. Tout ou partie de cette intelligence est la conséquence de leurs interactions (phénomène d'émergence). Un agent est une entité physique ou virtuelle (virtuelle dans notre cas) capable d'agir sur son environnement (les autres agents), qui peut communiquer directement avec d'autres agents, qui possède des ressources propres, qui est capable de percevoir son environnement, qui possède des compétences et offre des services (Ferber, 1995). Les SMA sont utilisés depuis longtemps dans le domaine des langues naturelles. Dès le début des années 70, des travaux en IAD ont été effectués sur la compréhension automatique de la parole (HEARSAY-II (Erman *et al.*, 1980)). Plus récemment, Lebarbé (2001) utilise cette approche pour l'analyse syntaxique tandis que Menézo *et al.* (1996) l'utilisent pour la détection d'erreurs. Parmi les autres travaux, certains se rapprochent de ce que nous voulons faire, une architecture modulaire permettant l'apprentissage de données et leur utilisation. On peut citer le système CARAMEL (Sabah, 1990) ou le système TALISMAN (Stéphanini *et al.*, 1992).

Les systèmes SMA offrent deux avantages techniques importants qui nous ont conduit a choisir ce mode d'implémentation. (1) la *possibilité de distribuer sur plusieurs machines* : les systèmes TALN ont toujours été consommateurs de ressources systèmes importantes dues aux données à stocker (la taille du lexique est d'au moins 100000 entrées pour une langue comme le Français) et aux calculs souvent lourds. Chaque agent peut se trouver sur une machine. Ainsi, il pourra pleinement utiliser les ressources matérielles disponibles. La communication entre agents se fait alors par accès réseaux sur le modèle client-serveur. (2) La *modularité* qui est une des caractéristiques principales des architectures agent. Un service global résulte de la coopération de chaque agent qui effectue une sous-tâche moins complexe. Les avantages génie-logiciel sont nombreux : un développeur peut facilement créer un agent dans le langage informatique de son choix, pourra aisément le tester et l'améliorer indépendamment des autres. L'ajout d'un programme déjà existant est simplifié. Dans notre application, par exemple, il a été extrêmement

facile de créer un agent d'analyse morpho-syntaxique qui n'est qu'une interface à l'analyseur SYGMART (Chauché, 1984).

4.2. *Exemple de société d'agents : implémentation*

4.2.1. *description du prototype*

Nous avons donc réalisé un prototype[5] sous la forme d'un système multi-agents. Chaque agent possède un certain nombre de compétences, une mémoire qui lui est propre et peut interagir avec son environnement (les autres agents). La gestion des agents se fait par un référenceur. Lors de sa création, chaque agent adresse au référenceur son identifiant, son rôle, éventuellement sa langue[6], ainsi que la machine et le port sur lequel il écoute. Le référenceur accepte la création de l'agent si aucun autre agent encore actif ne présente cet identifiant. Plusieurs types de communications sont possibles : (1) Communication directe entre agents : comme pour le *pair à pair* (*peer to peer*), un agent demande au référenceur l'adresse particulière d'un agent puis communique directement avec lui. C'est le cas, par exemple, d'un agent *contextualiseur* qui a besoin de récupérer très souvent les informations sur les lexies (cf. 4.2.2.) que stocke la *base*. (2) Communication par l'intermédiaire du référenceur : suivant la requête, les messages peuvent être envoyés à un agent, à tous les agents ayant un certain rôle, à tous les agents d'une langue ayant un certain rôle et même à tous les agents. En pratique, le message est envoyé au référenceur qui se charge de l'envoyer à ses destinataires. Chaque agent qui a reçu le message y répond en apportant une réponse ou en signifiant sa non-compétence dans ce domaine.

4.2.2. *Agents implémentés*

En Octobre 2003, plus d'une dizaine de types d'agents étaient implémentés.

- *Base lexicale* : ces agents implémentent l'architecture présentée en 3.4.
- *Contextualiseur* : cette sorte d'agents est capable de calculer le vecteur conceptuel d'un item en fonction de contextes sémantiques et morphologiques. En pratique, il fait une somme pondérée des vecteurs des *acceptions* du terme en fonction d'un vecteur contexte, d'informations morphologiques et statistiques (fréquence en corpus).
- *Analyseur morpho-syntaxique* : il s'agit de l'analyseur SYGMART (Chauché, 1984). Pour un texte, il renvoie l'arbre morpho-syntaxique correspondant.
- *Analyseur sémantique* : avec l'aide de l'agent d'analyse morpho-syntaxique et l'aide de l'agent contextualiseur, l'agent d'analyse sémantique calcule le vecteur correspondant à un texte. Ce texte, dans le cas d'un apprentissage, est une définition de dictionnaire dont l'agent d'apprentissage souhaite le vecteur.
- *Catégoriseur* (Jalabert, 2003) qui catégorise les *lexies* pour fabriquer les *acceptions*.
- *Apprentissage* : Cet agent gère l'apprentissage des lexies. Il est directement aidé dans cette tâche par des agents d'analyse sémantique ainsi que par les agents extracteurs de définitions.
- *Agents extracteurs de définitions* : ces agents ont pour rôle de récupérer les définitions correspondant à des items et de les fournir à l'agent d'apprentissage.
- *Agents calculant les fonctions lexicales* : ces agents sont des experts des relations sémantiques comme la synonymie, l'antonymie, l'hypéronymie ou toute autre fonction lexicale.
- *Agents extrayant les fonctions lexicales* : ces agents extraient des dictionnaires spécialisés ou de définitions de dictionnaires classiques des relations sémantiques qui existent entre les items.

[5] Accessible en ligne à l'adresse http//www.lirmm.fr/~schwab.

[6] Certains agents peuvent être indépendants de la langue. C'est le cas, par exemple, d'agents dont les données ne sont pas de type lexical.

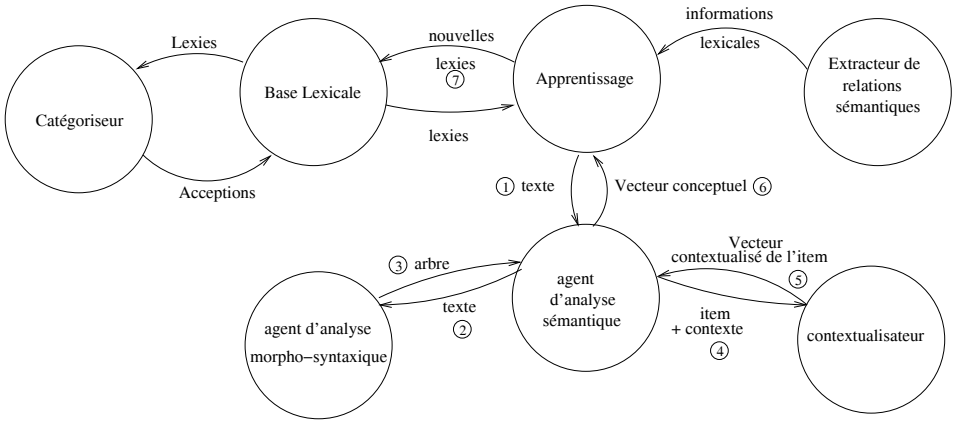


Figure 4. Organisation macroscopique du système au cours d'une analyse sémantique.

4.2.3. *Exemple d'interaction entre agents, apprentissage du vecteur d'un item*

Dans cette section, nous illustrons la collaboration entre agents par l'exemple de la fabrication d'objets *acceptions* et *lexies* d'un item lexical. L'agent d'*apprentissage* demande à un agent *extracteur de définitions* de lui fournir les définitions de cet item qu'il a récupérées en explorant le web ou des dictionnaires à usage humain. Le texte de chaque définition est donné à l'agent d'*analyse sémantique* (1) qui, à partir de l'arbre morpho-syntaxique obtenu grâce à l'agent *analyseur morpho-syntaxique* (2)-(3), calcule le vecteur conceptuel correspondant à la définition (6). L'agent *contextualiseur* (lui-même aidé par la *base de vecteurs*) et éventuellement des agents experts en *relations sémantiques* collaborent avec lui dans cette tâche (4)-(5). L'agent *apprentissage* récupère chaque vecteur des définitions. Les *agents extrayant les fonctions lexicales* lui permettent de compléter leurs informations pour construire les *lexies* qu'ils fournissent alors à la *base lexicale* (7). Parallèlement à ces opérations, l'agent *catégoriseur* explore la base et fabrique des *acceptions* à partir des *lexies* (8).

5. Conclusion et perspectives

Dans cet article, nous présentons l'architecture d'une base lexicale sémantique générique utilisable par des applications de TALN nécessitant l'usage de la sémantique comme, par exemple, la recherche d'informations, la traduction automatique ou le résumé automatique. À chaque item lexical sont associées des acceptions qui rassemblent les informations correspondants aux différents sens que l'item peut prendre. Ces acceptions sont fabriquées automatiquement à partir de définitions issues de différentes sources. Cette architecture se fonde sur plusieurs hypothèses. Nous avons montré que nous devions représenter le sens par une approche combinant approche thématique (par vecteurs conceptuels) et approche lexicale (par une explicitation des relations sémantiques externes) afin de tenir compte des informations spécifiques sur les rapports de sens que les termes entretiennent entre eux. De même, il est nécessaire de tenir compte des relations sémantiques internes (polysémie). Il convient aussi de réaliser une analyse multi-sources afin d'obtenir un maximum d'informations sur le lexique pour ne pas passer à coté d'informations qui pourraient se révéler importantes ainsi qu'un apprentissage permanent pour pouvoir tenir compte des informations extraites par certaines sources dynamiques comme le web. Nous avons finalement présenté certains agents qui ont été implémentés et qui sont accessibles en ligne.

Dans la suite de nos travaux, nous allons continuer à travailler sur chaque agent afin d'améliorer leurs tâches respectives. De nouveaux seront ajoutés essentiellement en ce qui concerne d'autres fonctions lexicales comme les fonctions hiérarchiques (hypéronymie, holonymie,. . .). Nous allons aussi travailler sur l'amélioration des modes de communication entre agents en analysant,

par exemple, les moyens techniques qui permettraient aux agents de seconder un autre agent sans que celui-ci ait à demander de l'aide.

Références

Chauché J. (1990). Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TAL Information* : 17-24.

Chauché J. (1984). Un outil multidimensionnel de l'analyse du discours. In *Proceedings of Coling'84* : 11-15.

Deerwester S.C., Dumais S.T., Landauer T.K., Furnas G.W. et Harshman R.A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, vol. (41/6) : 391-407.

Erman L., Hayes-Roth F., Lesser V. et Reddy D. (1980). The hearsay-ii speech understanding system : Integration knowledge to resolve uncertainly. In *ACM Computing Surveys*, vol. (12).

Ferber J. (1995). *Les systèmes multi-agents. Vers une intelligence collective*. InterEditions.

Hachette (2000). *Le Nouveau Petit Robert. Dictionnaire alphabétique et analogique de la langue française*. Hachette.

Jalabert F. (2003). Catégorisation de définitions et nommage de sens. Mémoire de DEA. Université Montpellier II, LIRMM.

Kawamoto A.H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity : A parallel distributed processing account. *Journal of Memory and Language*, vol. (32) : 474-516.

Lafourcade M. et Prince V. (2001). Synonymies et vecteurs conceptuels. In *Actes de TALN'2001*.

Larousse (1992). *Thésaurus Larousse – Des idées aux mots, des mots aux idées*. Larousse.

Larousse (2001). *Le Petit Larousse Illustré 2001*. Larousse.

Lebarbé T. (2001). Vers une plate-forme multi-agents pour l'exploration et le traitement linguistique. In *Actes de TALN'2001*.

Menézo J., Genthial D. et Courtin J. (1996). Reconnaissances pluri-lexicales dans celine, un système multi-agents de detection et correction des erreurs. In *NLP+IA96 : International Conference on Natural Language Processing and Industrial Applications.* : 174-180.

Morin E. (1999). Extraction de liens sémantiques entre termes à partir de corpus techniques. Thèse de doctorat. Université de Nantes.

Sabah G. (1990). Caramel : Un système multi-expert pour le traitement automatique des langues. *Modèles linguistiques*, tome (XII/1).

Salton G. (1968). *Automatic Information Organisation and Retrieval*. McGrawHill.

Schwab D., Lafourcade M. et Prince V. (2002). Vers l'apprentissage automatique, pour et par les vecteurs conceptuels, de fonctions lexicales. L'exemple de l'antonymie. In *Actes de TALN 2002*, vol. (1).

Stéfanini M.-H., Berrendonner A., Lallich G. et Oquendo F. (1992). Talisman : un système multi-agents gouverné par des lois linguistiques pour le traitement de la langue naturelle. In *Proceedings of Coling 1992*.

Sérasset G. et Mangeot M. (2001). Papillon lexical databases project : monolingual dictionaries and interlingual links. In *NLPRS 2001* : 119-125.

Le rappresentazioni di "Genere" e il mondo del lavoro attraverso la comunicazione scientifica in psicologia sociale: analisi del lessico degli "*Psychological Abstracts*" (1976-2002)

Gilda Sensales, Antonio Chirumbolo

Dipartimento di Psicologia dei Processi di Sviluppo e Socializzazione
Università "La Sapienza" – Via dei Marsi, 78 – 00185 Roma – Italia
chirumbolo@uniroma1.it

Abstract

On the one hand the research can be framed within the tradition of studies on Social Representations (SR) with a particular focus on the relationship between SR and scientific knowledge; on the other hand, it can be placed within the recent development of lexical analysis conducted on titles, abstracts or key-words of scientific articles. In our case the analysis is conducted on the entire corpus of "Psychological Abstracts". The aim of this investigation is the exploration of the lexical-semantic structure of the representations of the "gender and the world of work" within the scientific discourse of Social Psychology. This structure was also studied as regards to possible differential aspects. The research sample is composed of 162 abstracts selected by means of an automatic research on the PsychoLIT, the electronic database of the "Psychological Abstracts", through the following key-words: "GENDER AND SOCIAL PSYCHOLOGY AND WORK". The selected abstracts range from the year 1976 to the first semester of 2002. The structural characteristics were investigated with the ASPAR procedure of SPAD-T. The words of the abstracts, including titles, summaries, key-words, major description and classification codes, were considered as active variables. They were differentiated from each other by means of the inclusion of a specific letter in front of the lexical unit of each field (e.g. K in front of the word referring to the key-word). Data referring to, for instance, "country in which the journal was published", "gender of the first and of the second author", "period of publication", were considered as illustrative variables. After a preliminary phase of "data screening" with the EQUIVALENCE procedure of the SPAD-T, words with a frequency higher than 5 were taken into consideration and the analysis was run on 13,765 words, 606 of which were different (4.4%). The analysis pointed out three factorial axes clearly different from each other. Their valence is also discussed as regards the illustrative variables and the representational process on the scientific discourse and on common sense.

Riassunto

La ricerca è inquadrabile, da una parte nella tradizione degli studi sulle rappresentazioni sociali (RS), con particolare attenzione per il rapporto tra RS e conoscenza scientifica, dall'altra negli sviluppi più recenti di analisi lessicali condotte sui titoli o sui riassunti e/o le parole-chiave di articoli scientifici. Nel nostro caso l'analisi è condotta sugli "*Psychological Abstracts*". Obiettivo dell'indagine è l'esplorazione della strutturazione semantico-lessicale delle rappresentazioni di "genere" e "mondo del lavoro" nel discorso scientifico psicologico-sociale. Tale strutturazione è studiata anche in relazione agli eventuali aspetti differenziali legati alla sua articolazione. La popolazione della ricerca è composta di 162 abstracts individuati con una ricerca automatizzata sullo PsychoLIT, il database computerizzato degli "*Psychological Abstracts*", attraverso le parole chiave: "*GENDER AND SOCIAL PSYCHOLOGY AND WORK*". Essi coprono un arco di tempo che va dal 1976 al I semestre del 2002. Gli aspetti strutturali sono stati analizzati con la tappa ASPAR dello SPAD-T. Sono state considerate come variabili attive le parole degli interi abstracts comprensivi di titoli, riassunti, parole-chiave, descrizioni maggiori e codici classificatori. Esse sono state differenziate le une dalle altre attraverso l'introduzione di una specifica lettera davanti all'unità lessicale relativa a ciascun campo (ad esempio K davanti al termine riferito alle parole-chiave). Le variabili illustrative sono invece rappresentate dai dati categoriali relativi ad esempio al "paese di

pubblicazione della rivista", al "sesso del primo e del secondo autore", al "periodo di pubblicazione". Dopo una fase iniziale di "pulizia dei dati" con la tappa "equivalence" dello SPAD-T si è fissata la soglia ≥ 5 che ha permesso di condurre l'analisi su 13.765 parole, di cui 606 diverse, pari al 4.4% del totale. L'analisi ha evidenziato tre assi fattoriali nettamente distinti fra di loro. La loro valenza è discussa anche in relazione alle variabili illustrative ed agli elementi di continuità tra processi rappresentazionali che agiscono nel discorso scientifico e nel senso comune. [1]

Parole-chiave: genere, mondo del lavoro, psicologia sociale, rappresentazioni sociali, rassegna della letteratura, analisi lessicale, analisi delle corrispondenze.

Keywords: gender, work, social psychology, social representations, literature review, lexical analisys, correspondence analysis.

1. Introduzione

Lo studio si presenta come una forma particolare di rassegna della letteratura di cui si trovano esempi nei lavori di Doise, Staerkle, Clémence (1994), di Courtial (1999) e di Trombetta, Chirumbolo, Bruni (2001). Essi si riferiscono ad analisi testuali dei titoli o delle parole-chiave e/o degli interi abstracts condotte, nei primi due casi, utilizzando il database elettronico dello Psych-LIT, l'archivio elettronico su CD degli *Psychological Abstracts* curati dall'*American Psychological Association*; nel terzo caso, facendo riferimento ad articoli scientifici individuati attraverso lo spoglio manuale di alcune delle maggiori riviste psico-pedagogiche italiane dal 1960 al 1990.

Nell'insieme tali lavori rappresentano il tentativo di delineare, attraverso elaborazioni statistiche di dati testuali, la strutturazione ed evoluzione dei principali temi della ricerca scientifica legata a campi determinati della psicologia. Si tratta dunque di un tipo di rassegna della letteratura non ortodossa che, diversamente da quelle classiche di tipo narrativo, condivide piuttosto la "filosofia" obiettivista insita negli studi meta-analitici. Questi ultimi, definiti nei termini di integrazione statistica dei risultati di studi indipendenti, a partire dalla seconda metà degli anni '70, si sono andati affermando in campo psicologico (cf. Glass, 1976 ; Hedges and Olkin, 1985 ; Mullen, 1989 ; Rosenthal, 1991) per la possibilità di far raggiungere, come maggiore beneficio, un accumulo di conoscenza, quando non addirittura di portare a scoperte scientifiche (cf. in proposito il punto di vista di Sohn, 1996 che, pur favorevole alla tecnica meta-analitica, contesta quest'ultima possibilità).

Non è questa la sede per riprendere l'ampio dibattito che in psicologia si è venuto sviluppando proprio sulla plausibilità di tali tecniche (cf. in proposito Gergen e Gergen, 1984 ; Hunter e Schmidt, 1990 ; Miller e Cooper, 1991 ; Mullen *et al.*, 2001 ; Rosental e Di Matteo, 2001). Qui basterà ricordare come meta-analisi e rassegne della letteratura, al di là delle diverse opzioni epistemologiche ad esse sottese, vengano utilizzate per permettere di delineare punti di forza e di debolezza presenti nel corpus disciplinare consentendo di superare un limite di fondo della psicologia, legato ad un modo di fare scienza in cui ha operato il cosidetto "mito del '*lonely paradigm*'" (Moscovici, 1984) che ha avuto come conseguenza la tendenza ad ignorare la propria storia favorendo il susseguirsi di paradigmi di ricerca, senza un confronto con i paradigmi similari contemporanei o antecedenti.

In questo quadro uno dei primi tentativi di fare i conti sia con un bisogno "forte" di teorizzazione, sia con i criteri di "obiettività", propri dei metodi meta-analitici, in cui il momento interpretativo è limitato alla fase finale di discussione dei risultati, è riconducibile alla riflessione meta-teorica avviata dagli inizi degli anni '90 da de Rosa attraverso la messa a punto di

[1] Il lavoro è stato condotto basandosi sui dati elaborati nella tesi di laurea di Luciana Preite, della quale è stata relatrice Gilda Sensales, mentre Antonio Chirumbolo ha svolto la supervisionie delle elaborazioni statistiche.

una griglia di analisi del contenuto da applicarsi alla letteratura relativa agli studi sulle rappresentazioni sociali (cf. de Rosa, 1994 e 2001 ; de Rosa e D'Ambrosio, 2003).

La particolare rassegna della letteratura da noi condotta, condivide le esigenze appena ricordate e, rispetto ai lavori citati in apertura, si presenta come una proposta innovativa perché in grado di prendere in considerazione contemporaneamente il titolo e praticamente quasi l'intero corpus degli abstracts attraverso degli stratagemmi operazionali che saranno illustrati nella sezione metodologica (cf. anche Sensales, in stampa).

Per quello che riguarda poi l'inquadramento teorico generale dello studio, esso è riconducibile alla tradizione delle rappresentazioni sociali (RS) in riferimento soprattutto al ruolo da esse giocato nel discorso scientifico. E' questo un tema relativamente poco trattato poiché generalmente gli studi interni a questa prospettiva hanno privilegiato l'analisi delle RS rispecchiate e costruite dal senso comune, benché da parte di alcuni sia stata affermata con forza l'importanza dell'esplorazione del rapporto fra RS e sapere specialistico (Sensales, 1990 e 1994 ; Courtial, 1999).

Infine a livello di contenuti l'argomento trattato è riferito al tema del "genere" nel mondo del lavoro, esplorato in una prospettiva psicologico-sociale.

Tale tema rappresenta un terreno di studio che si deve confrontare con due realtà particolarmente in evoluzione: quella legata all' "identità di genere" (cf. Gelli, 2002) e quella del "mercato del lavoro" a proposito delle quali si fa riferimento a meccanismi di "ibridazione" che opererebbero al loro interno stimolando profonde trasformazioni (cf. Walkerdine *et al.*, 2001).

Da questo punto di vista, dunque, appare di particolare interesse poter analizzare il lessico di cui si compone la comunicazione scientifica al fine di individuare temi specifici che hanno attraversato la conoscenza psicologico-sociale caratterizzandone, in ipotesi, determinati periodi e contesti socio-culturali.

Obiettivo della ricerca

La ricerca, a carattere esploratorio, è condotta sugli "*Psychological Abstracts*" con l'obiettivo di studiare la strutturazione semantico-lessicale, delle rappresentazioni di "genere e mondo del lavoro" nel discorso scientifico psicologico-sociale. Tale strutturazione sarà studiata anche rispetto agli eventuali aspetti differenziali legati alla sua articolazione.

2. Metodologia

2.1. *Popolazione della ricerca*

La nostra popolazione è composta di 162 abstracts individuati con una ricerca automatizzata sullo PsychoLIT, il database computerizzato degli "*Psychological Abstracts*", attraverso le parole chiave: "*GENDER AND SOCIAL PSYCHOLOGY AND WORK*", selezionando un arco di tempo che va dal 1887 al primo semestre del 2002. La ricerca non ha dato risultati fino al 1976[2]. Da quell'anno sono apparsi i primi abstracts così distribuiti:

[2] In proposito va ricordato come per la prima volta nel 1974 vi fu, contemporaneamente da parte di psicologi diversi, una riconcettualizzazione delle categorie relative ai ruoli sessuali in termini di "*gender*" (cf. Pearson e Cooks, 1995 ; cf. anche Martin, 1995).

Periodo	*Frequenza Abstracts*
1976-1987	23
1988-1993	24
1994-1996	34
1997-1999	36
2000-2002	45

2.2. *Percorso operativo*

Una volta individuati gli Abstracts essi sono stati importati in un apposito file di testo, mentre si è proceduto parallelamente ad applicare una griglia di analisi di contentuo in grado di rilevare informazioni relative ad esempio al "paese di pubblicazione della rivista", al "sesso del primo e del secondo autore", alla "tipologia dell'articolo", al "periodo di pubblicazione", al "titolo della rivista", trattate come variabili categoriali inserite in un apposito file numerico.

2.3. *Elaborazioni statistiche*

Sugli Abstracts è stata condotta una analisi lessicale con il pacchetto statistico SPAD-T (Lebart *et al.*, 1989) utilizzando come corpus d'analisi le parole dell'intero abstract comprensivo del "titolo", del "riassunto dell'articolo", delle "parole chiave", delle "descrizioni maggiori", dei "codici classificatori". Negli ultimi tre casi quando si trattava di parole composte si sono riportate senza lo spazio per evitare di perdere il loro senso compiuto.

Questi primi cinque campi rappresentano l'unità di contesto, mentre l'unità di testo (su questa distinzione cf. Bolasco, 1999) è la forma grafica, l'unità lessicale, presente nei cinque campi e differenziata rispetto ai campi attraverso l'introduzione della lettera T, K, M, C, rispettivamente davanti alle parole del titolo, delle parole-chiave, delle descrizioni maggiori, del codice classificatore, mentre l'assenza di una lettera maiuscola all'inizio della parola indica la sua appartenenza al riassunto vero e proprio.

Gli altri campi, rilevati attraverso la griglia di analisi del contenuto, sono considerati variabili categoriali trasferite, come si è detto, in un file numerico a parte. Qui basterà ricordare che tali variabili sono 14 per un totale di 22 modalità considerate.

Le unità statistiche di analisi (frammenti o osservazioni) prese in considerazione sono gli abstracts comprensivi di titoli, riassunti, parole-chiave, descrizioni maggiori e codici classificatori, mentre le unità lessicali (forme) sono le parole contenute nei cinque campi. La matrice dei dati è costituita dai "frammenti" x le "forme" (cf. Bolasco, 1999 ; Trombetta *et al.*, 2001), in cui in riga si hanno i frammenti di testo, cioè gli abstracts, e in colonna si hanno le forme, cioè le parole. A questa matrice si sono associati, per ogni abstract, le variabili categoriali sopra citate.

Per analizzare la matrice si è applicata l'analisi delle corrispondenze per dati testuali (AC), utilizzando, come si è già ricordato, il pacchetto statistico SPAD-T.

In particolare sono state condotte diverse analisi lessicali con tappe differenziate dello SPAD-T:

una prima tappa dello SPAD-T ha permesso di evidenziare il dizionario di tutte le parole;

la successiva tappa delle "*equivalences*" ha reso possibile intervenire sul dizionario per accorpare le forme contigue quali ad esempio singolari e plurali;

una analisi di tipo strutturale, con la procedura ASPAR, per l'AC, ha consentito di individuare le principali dimensioni latenti sottostanti il lessico degli abstracts. Esse sono definite in termini di fattori e sono atte a riassumere l'intreccio delle relazioni di interdipendenza tra le

variabili originarie. L'analisi è condotta sulle unità lessicali – le parole degli interi abstracts comprensivi di titoli, riassunti, parole-chiave, descrizioni maggiori e codici classificatori - considerate come variabili attive, variabili cioè che concorrono direttamente all'individuazione degli assi fattoriali, e sui dati categoriali – come ad esempio i già citati "paese di pubblicazione della rivista", "sesso del primo e del secondo autore", "tipologia dell'articolo", "periodo di pubblicazione", "titolo della rivista" – considerati come variabili illustrative, variabili proiettate sugli assi fattoriali per aiutarne l'interpretazione permettendo inoltre di individuare eventuali aspetti differenziali[3].

3. Risultati

Il file testo iniziale è composto di 15.510 parole, di cui 3.755 diverse, pari al 24.2% del totale. Si è avuta una fase preliminare di "pulizia dei dati" – con l'eliminazione di forme quali ad esempio gli articoli, le preposizioni, le congiunzioni – e successivamente si è applicata la tappa delle *equivalences* che ha permesso una aggregazione delle forme lessicali contigue quali ad esempio i singolari ed i plurali, i presenti, passati e infiniti dello stesso verbo, per un totale di 516 equivalenze. A causa della grande dispersione dei dati, con circa 1754 parole con soglia da 1 a 5, si è fissata la soglia ≥ 5. Si è così condotta l'analisi su 13.765 parole, di cui 606 diverse, pari al 4.4% del totale.

In base allo scree-test si è deciso di prendere in considerazione i primi tre fattori in grado di spiegare complessivamente il 6.04% della inerzia totale.

L'articolazione lessicale dei fattori è illustrata nelle apposite tabelle. Per ogni fattore sono state ritenute le parole con contributo assoluto (C.A.) rilevante. Poiché la somma di tutti i contributi assoluti è pari a 100, si sono considerate, per le variabili attive, le parole con C.A. ≥ .2 (si è applicata la formula 100/N-1, dove N è uguale al numero di parole distinte, che ha permesso di stabilire come rilevante il CA ≥ .16), mentre per le variabili illustrative si sono considerate le modalità con C.A. ≥ 4.76 (100/21 modalità = 4.76). Nelle tabelle sono riportate solo le prime 44 parole, di cui quelle riferite al titolo, alla parola-chiave, alla descrizione maggiore, al codice classificatore, sono evidenziate in grassetto.

Il I Fattore mostra (cf. Tab. 1), sul semiasse negativo, 41 parole, di cui tre riferite rispettivamente al termine "*Family*" presente nel titolo, alla parola-chiave "*Families*", al codice classificatore "*Marriage family*"; mentre sul semiasse positivo sono presenti 24 parole, di cui solo una è riferita alla parola-chiave "*College Student*". In considerazione dell'articolazione del lessico, questo I fattore è stato definito della "Comunicazione scientifica centrata sulle questioni di genere, da una parte, all'interno della famiglia, con una particolare attenzione per i ruoli familiari e gli aspetti di conflitto ed insoddisfazione, dall'altra, all'interno dell'ambiente scolastico con una focalizzazione sui processi di discriminazione".

Il II Fattore evidenzia (cf. Tab. 2), sul semiasse negativo, 58 parole, di cui sette fanno riferimento rispettivamente a tre termini appartenenti al titolo, "*Harassment*", "*Sexual*", "*Theory*", e a quattro parole-chiave, "*Health*", "*Sexual*", "*Harassment*", "*Social*"; mentre il semiasse positivo evidenzia 43, parole di cui 9 sono riferite rispettivamente a termini presenti nel titolo, "*Applicant*", "*Job*", "*Effect*", "*Gender*", alle parole-chiave, "*Applicant*", "*Job*", "*Hiring*", al codice classificatore "*Personnel management*", alla descrizione maggiore "*Job applicant*". In base al lessico che compone questo II fattore, esso è denominato della

[3] In altri termini, mentre la posizione delle parole su ciascun asse viene stabilita dalla loro reciproca co-occorrenza, determinando quindi attivamente gli assi, la posizione ad esempio delle riviste e degli anni di pubblicazione, rispetto agli assi, viene considerata dopo aver individuato i fattori, allo scopo di aiutare ad illustrare, descrivere e chiarire il loro significato.

Tabella 1. I Fattore (varianza spiegata: 2.16%)

SEMIASSE NEGATIVO		*SEMIASSE POSITIVO*	
VARIABILI ATTIVE	C.A.	VARIABILI ATTIVE	C.A.
Husband	7.9	Male	1.1
Spouse	7.0	Evaluate	.4
Wife	6.6	Undergraduate	.4
Withdraw	4.6	Discrimination	.3
Discontent	5.1	Feature	.3
Contribution	4.5	Small sample	.3
Conflict	3.0	Suitability	.3
Wives	2.5	Trait	.3
Demand	2.0	Aggressive	.2
Family	1.8	Bias	.2
Division	1.7	College	.2
Interaction	1.6	Decision	.2
Change	1.4	Gender	.2
Report	1.2	Group	.2
CMarriagefamily	1.1	**KCollegeStudent**	.2
Labor	1.1	Low	.2
Adress	1.0	Member	.2
Couple	.7	Outgroup	.2
Response	.7	Pay	.2
Marital	.5	Respect	.2
Responded	.5	Stereotype	.2
TFamily	.5	Student	.2
Increase	.4	Style	.2
KFamilies	.4	Varied	.2

Life	.4		
Work	.4		
Data	.3		
Distress	.3		
Domain	.3		
Gendered	.3		
Home	.3		
Quality	.3		
Well	.3		
Career	.2		
Child	.2		
Dual	.2		
During	.2		
Final	.2		
Pattern	.2		
Remain	.2		
While	.2		
VARIABILI ILLUSTRATIVE	V.TEST	VARIABILI ILLUSTRATIVE	V.TEST
Paese pubb.: Europa	-38.4	Rivista: J.Applied S.P.	20.6
Autore: Unico A. Donna	-14.1	Paese pubb.: USA	18.2
Periodo pub.1997-1999	-12.3	Paese pubb.: Oriente	16.5
		Autore: Tutte Donne	11.4
		Rivista: J.Social Psych.	8.3

Tabella 2. II Fattore (varianza spiegata: 2.06%)

SEMIASSE NEGATIVO		*SEMIASSE POSITIVO*	
VARIABILI ATTIVE	C.A.	VARIABILI ATTIVE	C.A.
Social	1.9	Applicant	2.7
Sexual	1.5	Job	1.4
Harassment	1.2	Likelihood	1.4
Health	1.0	Candidate	1.2
KHealth	1.0	Accent	.8
Cognitive	.9	Like	.8
Identy	.9	Subject	.7
Research	.8	Hire	.6
Model	.7	Name	.6
Psychology	.7	Sex	.6
Review	.7	**KApplicant**	.5
KSexual	.6	Scenarios	.5
Psychological	.6	Stereotypical	.5

Stress	.6	**CPersonnelManagem.**	.4
Theoretical	.6	Expected	.4
Class	.5	Information	.4
Concept	.5	**KJob**	.4
THarassment	.5	Manipulated	.4
TSexual	.5	Masculine	.4
Analyses	.4	Match	.4
Author	.4	Status	.4
Empirical	.4	**TApplicant**	.4
KHarassment	.4	Experiment	.3
Major	.4	**KHiring**	.3
Practice	.4	Label	.3
Context	.3	Masculinity	.3
Control	.3	**MJobApplicant**	.3
Difference	.3	Occupation	.3

Explore	.3	Require	.3
Implications	.3	Stereotypic	.3
Individual	.3	Task	.3
KSocial	.3	**TJob**	.3
Lived	.3	Expectations	.2
Mental	.3	Goal	.2
Multiple	.3	Greater	.2
Power	.3	Investigated	.2
Problem	.3	Mediate	.2
Professional	.3	Perceive	.2
TTheory	.3	Preference	.2
Educational	.2	Salary	.2

Development	.2	Success	.2
Coping	.2	**TEffect**	.2
Personality	.2	**TGender**	.2
Cultural	.2		
VARIABILI ILLUSTRATIVE	V. TEST	VARIABILI ILLUSTRATIVE	V. TEST
Periodo pub.2000-2002	-17.3	Autore: Primo A. Donna	19.9
Autore: Unico A. Uomo	-16.3	Periodo pub.1976-1987	15.2
Periodo pub.1994-1996	-9.7	Rivista: S.P. Quarterly	9.7
Paese pubb.: Canada	-5.4		

Tabella 3. III Fattore (varianza spiegata: 1.82%)

SEMIASSE NEGATIVO		*SEMIASSE POSITIVO*	
VARIABILI ATTIVE	C.A.	VARIABILI ATTIVE	C.A.
White	.5	Seeker	18.6
Black	.3	Reassurance	11.9
Against	.2	Depression	10.4
Attitude	.2	Feedback	9.1
Political	.2	Integrated	5.4
Solo	.2	Symptoms	4.2
Toward	.2	Target	3.9
		Negative	3.1
		Time	2.8
		KTheory	2.0
		Interpresonal	2.0
		Theory	1.5
		Engaged	1.1
		Self	1.0
		Appraisal	.9
		Late	.7

		Assess	.6
		Consistent	.5
		Enhance	.4
		Evaluation	.4
		Risk	.3
		Closed	.3
		Applied	.3
		Violate	.2
		Test	.2
		Place	.2
		Combine	.2
VARIABILI ILLUSTRATIVE	V. TEST	VARIABILI ILLUSTRATIVE	V. TEST
		Periodo pub.1988-1993	19.2
		Autore: Primo A. Uomo	18.9
		Rivista: J. Person. S.P.	15.2

"Comunicazione scientifica orientata, da una parte, sulla salute e sulle molestie sessuali ma anche su fonti di diseguaglianza e oppressione diverse da quelle di genere, dall'altra, sugli aspetti occupazionali veri e propri con una particolare attenzione per i processi di stereotipizzazione e per le differenze di genere anche focalizzate sui tratti maschili".

Il III Fattore presenta (cf. Tab. 3), sul semiasse negativo, solo 7 parole, tutte parti del riassunto vero e proprio, come ad esempio "*White*", "*Black*", "*Attitude*", o "*Political*"; mentre sul semiasse positivo si trovano 27 parole, di cui una rappresentata dalla parola-chiave "*Theory*" e le altre riferite al corpo del riassunto come ad esempio "*Reassurance*", "*Depression*", o "*Symptoms*". In riferimento al lessico di questo III fattore, esso può essere descritto come della "Comunicazione scientifica articolata, da una parte, su temi etnici e politici, e dall'altra su aspetti, anche caratterizzati teoricamente, legati alla salute psichica in generale".

Per ciascuno dei tre fattori le tabelle indicano anche la presenza significativa di alcune delle variabili illustrative. Il loro posizionamento sarà commentato qui di seguito nella sezione delle conclusioni.

4. Conclusioni

Nell'insieme le rappresentazioni dei problemi di "genere" nel mondo del lavoro, veicolate dalla comunicazione scientifica di ambito psicologico-sociale, appaiono articolate in tre grandi dimensioni ciascuna bipolare e caratterizzata da una presenza diversificata di quelle che, nell'analisi, abbiamo definito variabili illustrative.

Per quello che riguarda la comunicazione scientifica focalizzata sul tema del "genere" in ambito domestico o scolastico (prima dimensioine), si può notare come il lessico degli abstracts è organizzato intorno a due argomenti posizionati lungo un ideale continuum e centrali rispetto all'oggetto di indagine. Essi sono legati alla visione più tradizionale del ruolo femminile: uno relativo al lavoro domestico e l'altro riferito al lavoro in campo educativo (cf. Fassinger, 2001). Entrambi gli ambiti sono marcati da elementi conflittuali: o riferiti esplicitamente alla "insoddisfazione", al "conflitto", all'"angoscia" legati alle mansioni domestiche, o indirettamente richiamati dai termini "discriminazione", "aggressivo", "stereotipo", rapportati al lavoro in ambiente scolastico.

Così, nel primo caso, gli elementi di tensione potrebbero alludere al disagio per attività onerose ma non retribuite, quali quelle domestiche, che delineano una sfera separata da quella del lavoro propriamente detto portando a una "doppia presenza" della donna contrassegnata, secondo la letteratura specialistica, dal "conflitto lavoro/famiglia" in cui si fronteggerebbero, da una parte la cura della casa e dei figli e, dall'altra, la voglia/necessità di proiettarsi all'esterno con un conseguente riconoscimento in termini monetari e di carriera (Barnett, 2001). Da questo punto di vista i temi evocati dal lessico di questa polarità sono riconducibili teoricamente ad una prospettiva psicologica di "scarsità" piuttosto che di "espansione" delle energie. Secondo questa interpretazione spendere energie in un campo comporterebbe necessariamente una diminuzione di energie spendibili in altri settori (Barnett, 2001). La presenza di elementi lessicali riferiti al disagio conferma questa lettura che è decisamente opposta ad una considerazione positiva dell'assunzione di ruoli molteplici, vista appunto come portatrice anche di benefici e di potenziamento della persona.

Qui val la pena ricordare come questi aspetti, che la nostra analisi riconduce alle ricerche pubblicate nel trienno '97-99, appaiano in realtà superati dalle più recenti tendenze di studio che sono invece focalizzate su un modello di integrazione fra esigenze lavorative e familiari proponendo un "bilanciamento lavoro/famiglia" non marcato dalle differenze di genere e quindi da una netta divisione dei ruoli (cf. Barnett, 2001). In realtà di questo modello non vi è traccia nei nostri dati plausibilmente perché è ancora un modello comportamentale e di ricerca ampiamente minoritario e quindi non rilevabile da una indagine come quella qui proposta; mentre altri tipi di elaborazioni, quali ad esempio l'analisi del lessico "tipico" dell'ultimo periodo da noi rilevato (tappa MOCAR dello SPAD-T) potrebbero, in ipotesi, permettere una verifica della sua presenza.

Per restare comunque ancora un momento alla prima dimensione si può notare come l'altra polarità, forse non casualmente volta verso la sfera educativa, considerata luogo privilegiato della segregazione orizzontale delle donne lavoratrici (cf. Piccone Stella e Saraceno, 1996), sia caratterizzata da un lessico che richiama la doppia discriminazione: quella di un lavoro che trasferisce la cura dei figli dall'area del privato al settore pubblico, con una convergenza fra

ruolo familiare e ruolo lavorativo, una professionalizzazione di attività in passato svolte esclusivamente nella sfera privata – non è un caso che, come si è appena ricordato, l'insegnamento sia per definizione considerata un'occupazione tipicamente femminile (anche se solo a livello di "divisione orizzontale del lavoro") –, e quella a cui si è educati attraverso la più importante agenzia di socializzazione rappresentata, dopo la famiglia, proprio dalla scuola (cf. Burr, 1998). Quindi l'istituzione scolastica quale possibile contesto in cui agisce la discriminazione a livello lavorativo e a livello di costruzione identitaria ma anche, e non da ultimo, come contesto della ricerca. E proprio a questo proposito si può citare la diffusa critica verso la sovra-utilizzazione della realtà universitaria e dei suoi soggetti quale riferimento per le indagini psicologiche che finirebbero in questo modo per essere autoreferenziali e scarsamente generalizzabili.

Passando alla comunicazione scientifica orientata prevalentemente sulla salute e sulle molestie sessuali o sugli aspetti occupazionali in senso stretto, si può per prima cosa notare come, a differenza della precedente dimensione, questa seconda mostri una cesura fra i due poli, segnalata anche dalla diversa scansione temporale in essi presente, che è in qualche modo sottolineata dalla caratterizzazione sociocentrica di un polo e da quella più sperimentalista del versante opposto. Da questo punto di vista il risultato appare particolarmente saliente per il semiasse riferito agli stress post-traumatici causati dalle molestie sessuali in cui sono chiamate in causa fonti di diseguaglianza ed oppressione evocate dalle parole "razza", "classe", "potere", probabilmente studiate in quanto partecipi del contesto in cui le molestie hanno luogo; mentre è meno scontato per il semiasse caratterizzato dai temi occupazionali, che sono trattati anche con un'operazionalizzazione in senso sperimentale e con una attenzione particolare per costrutti apparentemente individuocentrici, quali quelli legati ai tratti psicologici e di personalità implicati nelle categorie "mascolino/mascolinità". A tal proposito va però ricordato che il riferimento alla mascolinità è strettamente collegato al concetto stesso di genere, inteso come termine binario che implica un'attenzione per il rapporto di scambio e relazione continui tra i due sessi (cf. Ruspini, 2003). Questa seconda interpretazione, focalizzata anche sulla dimensione relazionale, sembra essere più coerente con la presenza, a livello delle cosiddette variabili illustrative proiettate su questo versante della dimensione, di una delle riviste più attente al contesto sociale delle relazioni quale il "*Social Psychology Quarterly*".

Esplicitando poi il riferimento alla diversa periodizzazione, si può sottolineare come gli studi centrati sulla salute e sulle molestie sessuali siano stati pubblicati soprattutto tra il 1994-1996 e tra il 2000-2002, confermando l'importanza e attualità del tema anche per la psicologia sociale. In proposito si può ricordare con Fitzgerald, Collinsworth, Harned (2001) come la ricerca in questo campo fosse sconosciuta fino a circa una dozzina di anni fa presentandosi adesso quale campo consistente e ben definito della psicologia con più di 350 articoli pubblicati, tra il 1996 ed il 2001, su riviste accreditate. Per contro, e possiamo dire non a caso, gli studi sugli aspetti occupazionali veri e propri sembrano segnare la nascita dell'interesse della psicologia sociale per il tema delle differenze di genere nel mondo del lavoro con una focalizzazione, però, sulle situazioni sperimentali in aperto contrasto con quelle critiche verso lo sperimentalismo che animavano il particolare periodo in cui erano pubblicati gli articoli (tra il 1976 ed il 1987). Esso corrisponde all'"era dei dubbi", alla "crisi di fiducia" che colpiva la disciplina attraverso la messa in discussione della sperimentazione quale caratteristica fino ad allora dominante nel campo (cf. Sensales, 2003). Ora però questa autoriflessione critica non sembra riguardare gli abstracts da noi indagati e questo probabilmente non tanto perché le autrici non siano inquadrabili in quel punto di vista definito di "empirismo femminista"

(Harding,1986), secondo il quale si utilizzavano strumentalmente metodologie positiviste per assicurarsi visibilità accademica, quanto piuttosto perché i lavori di cui ci stiamo occupando non sono riconducibili ad una prospettiva di studi propriamente femministi ma piuttosto a quelli sul genere in senso lato (cf. su questa differenza Pearson e Cooks, 1995). Essi sono infatti prodotti da ricercatori di entrambi i sessi, anche se con una studiosa come prima autrice, ed in questo senso sembrano seguire in modo aproblematico il solco della tradizione psicologica *mainstream* che privilegiava l'ottica sperimentalista su tutte le altre.

Riferendosi infine alla comunicazione scientifica articolata su temi etnici e politici o sulla salute psichica in generale (terza dimensione), si può notare come anch'essa presenti una netta separazione dei contenuti che danno vita ai due poli della dimensione di riferimento. In essi è riprodotto in forma amplificata il lessico del precedente semiasse negativo. Così da una parte ritorna l'interesse per le differenze etniche, ma con una esplicitazione della contrapposizione fra il gruppo orientato alla dominazione, i "bianchi", ed il gruppo centrato sulla subordinazione, i "neri"; mentre dall'altro si ripresenta il tema della salute indirettamente evocato da una terminologia di derivazione clinca anche orientata verso la patologia. Nell'insieme il lessico scientifico condensato in questa dimensione appare nettamente residuale rispetto al focus da noi indagato e purtuttavia consente una considerazione conclusiva che interessa l'articolazione complessiva delle rappresentazioni da noi investigate in relazione ad una specifica caratterizzazione: l'appartenenza di genere degli autori dei contributi repertoriati. Solo gli aspetti delle nostre rappresentazioni legati ad una situazione di difficoltà delle donne, sia in termini di salute e molestie sessuali sul luogo di lavoro, sia in termini di salute psichica in generale, sono trattati da ricercatori di genere maschile, mentre gli altri aspetti sono oggetto di una comunicazione scientifica svolta in prima persona da studiose. Ciò sembra alludere ad una sorta di doppio bias androcentrico (sui concetti di "gender bias" e di "selezione androcentrica dei temi di ricerca" in psicologia cf. Riger, 2002) all'opera in un sapere specialistico che seleziona temi legati alla patologia delle donne, in grado di garantire una rassicurante posizione gerarchicamente superiore dell'uomo che, come psicologo, le studia, e come studioso è unico estensore della pubblicazione o eventualmente affiancato, solo in seconda battuta, da una collega.

Questo risultato conclusivo potrà essere oggetto di approfondimenti successivi con elaborazioni in grado, in ipotesi, di evidenziare ulteriori aspetti differenziali legati alla diversa appartenenza di genere di coloro che hanno redatto gli articoli.

Infine due ultime considerazioni. La prima riguarda l'assenza di riferimenti a termini riconducibili in qualche modo a quei meccanismi di «ibridazione», ricordati nell'introduzione e relativi sia all'«identità di genere» che al «mercato del lavoro», la cui eventuale presenza potrà essere esplorata con specifiche tappe del programma statistico SPAD-T per l'individuazione di nuclei lessicali formati, ad esempio come è stato già ricordato, a partire dagli abstracts degli ultimi anni. I nostri risultati infatti possono essere stati ampiamente condizionati dal tipo di elaborazioni privilegiate che lavorano quantitativamente su dati qualitativi, ma in particolare tendono ad oscurare quegli aspetti di contenuto meno frequenti statisticamente, che potrebbero essere anche i più innovativi rispetto ai temi indagati. Da questo punto di vista altre tappe del programma potranno permettere di superare in parte tale limite, mentre resta auspicabile affiancare le tecniche quantitative con analisi anche di tipo narrativo in grado di recuperare quelle informazioni altrimenti ignorate.

La seconda annotazione concerne un bilancio complessivo sul lavoro da noi condotto che, da una parte, porta ad evidenziare la forte presenza di riferimenti ad elementi di disagio psicologico quale tratto caratterizzante il focus di studio delle questioni di genere in ambiente

lavorativo. Dall'altra, l'analisi della letteratura da noi repertoriata mostra una sorta di continuità fra sapere specialistico e senso comune, non solo perché entrambi i domini appaiono connotati da forme rappresentazionali semplificate, che, nel caso della conoscenza psicologico-sociale, hanno orientato l'ambito della ricerca riproponendo gerarchie consolidate, ma anche perché nell'insieme gli stessi temi che contrassegnano l'agenda del discorso scientifico segnalano lo stretto intreccio fra scienze sociali e arena pubblica. In proposito può valere per tutti l'esempio dell'attenzione per le molestie sessuali nei luoghi di lavoro che abbiamo visto essere un oggetto di studio centrale per la psicologia sociale in ambito lavorativo ma che sappiamo essere anche un argomento ampiamente trattato dai diversi media.

Con queste osservazioni si chiude per il momento il nostro lavoro che si è dimostrato in grado di stimolare una riflessione teorica di ampia portata ma anche strettamente vincolata ai dati empirici. Esso ha rivelato inoltre il valore critico della ricognizione della letteratura qui proposta che ha fornito una mappatura delle nostre conoscenze su uno specifico campo del sapere sottolineando, al tempo stesso, la continuità fra i processi psico-sociali di tipo rappresentazionale operanti nel discorso scientifico come in quello di senso comune, entrambi partecipi di pratiche socialmente costruite.

Bibliografia

Barnett R.C. (2001). Work-family balance. In Worell J. (ed.), *Encyclopedia of women and gender*. Academic Press.

Bolasco S. (1999). *Analisi mltidimensionale dei dati*. Carocci.

Burr V. (1998). *Gender and social psychology*. Routledge.

Courtial J.P. (1999). Analyse dynamique des representaions socilaes des chercheurs: le cas de l'autisme. *Papers on social representation*, vol. (8): 3.1-3.22.

Doise W., Staerkle e Clemence A. (1996). Vygotsky in the East and the West. A textual analysis of key-words and abstract in recent psychology journal (1974-1994). *Journal of Russian and East European Psychology*, vol. (34): 67-96.

Elms A.C. (1975). The crisis of confidence in social psychology. *American Psychologist*, vol. (30): 967-976.

Fassinger R. (2001). Women in non traditional work fields. In Worell J. (Ed.), *Encyclopedia of women and gender*. Academic Press.

Fitzgerald L., Collinsworth L.L. e Harned M. (2001). Sexual harassment. In Worell J. (Ed.), *Encyclopedia of women and gender*. Academic Press.

Gelli B.R. (Ed.) (2003). *Voci di donne*. Manni.

Gergen K.J. e Gergen M. (Eds) (1984). *Historical social psychology*. Lawrence Erlbaum.

Glass G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research*, vol. (5): 3-8.

Harding S. (1986). *The science question in feminism*. Cornell University Press.

Hedges L.V. e Olkin I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hunter J.E. e Schmidt F.L. (1990). *Methods of meta-analysis. correcting error and bias in research findings*. Sage.

Lebart L., Morineau A. e Bécue M. (1989). *SPAD T* . Cisia.

Martin C.L. (1995). Gender. In Manstead A.S.R. e Hewstone M. (Eds), *The Blackwell encyclopedia of social psychology*. Blackwell Publishers.

Miller N. e Cooper H.M. (eds) (1991). Meta-analysis in personality and social psychology. Special issue. *Personality and Social Psychology Bulletin*, vol. (17): 243-349.

Moscovici S. (1984). The myth of the lonely paradigm: a rejoinder. *Social Research*, vol. (51): 940-967.

Mullen B. (1989). *Advanced basic meta-analysis*. Lawrence Erlbaum.

Mullen B., Muellerleile P. e Bryant B. (2001). Cumulative meta-analysis: a consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin*, vol. (27): 1450-1462.

Pearson J.C. e Cooks L. (1995). Gender and power. In Kalbfleisch P.J. e Cody M.J. (Eds), *Gender, power, and communication in human relationships.* Lawrence Erlbaum Associates.

Piccone Stella S. e Saraceno C. (1996). Genere. *La costruzioine sociale del maschile e del femminile*. Il Mulino.

Riger S. (2002). Epistemological debates, feminist voices: sciences, social values, and the study of women. In Pickren W.E. e Dewsbury D.A. (Eds), *Evolving perspectives on the history of psychology*, American Psychological Association.

de Rosa A.M.S. (1994). From theory to metatheory in social representations: the lines of argument of a theoretical-methodological debate. *Social Science Information*, vol. (33): 273-304.

de Rosa A.M.S. (2001). Le besoin d'une théorie de la méthode. In Garnier C. (Ed.), *Les formes de la pensée sociale*. PUF.

de Rosa A.M.S. e D'Ambrosio M. (2003). An empirical example of the comparison between multiple correspondence analysis and smallest space analysis: the diffusion of the social representations theory through the institutional context of scientific communication. In Levy S. e Elizur D. (Eds), *Facet theory: towards cumulative social science*. Faculty of Arts, Center for Educatioin Development.

Rosenthal R. (1991). *Meta-analytic procedures for social research*. Sage.

Rosenthal R. e Di Matteo M.R. (2001). Meta-analysis: a consideration of indicators of sufficiency and stability. *Personality and Social Psychology Bulletin*, vol. (27): 1450-1462.

Ruspini E. (2003). *Le identità di genere*. Carocci.

Sensales G. (1990). *L'informatica nella stampa italiana*. Franco Angeli.

Sensales G. (1994). The communication systems of representations: a psychosocial research into representations of computers and informatics in italian daily newspapers. *Public Understanding of Science*, vol. (3): 347-363.

Sensales G. (2003). Percorsi teorico-critici in psicologia sociale: note per una ricognizione. In Sensales G. (Ed.), *Percorsi teorico-critici in psicologia sociale*. Franco Angeli.

Sensales G. (in stampa). Le radici della psicologia sociale attraverso la comunicazione scientifica internazionale: una analisi lessicale della "*Psychological Literature*" dal 1887 al 1954. In De Piccoli N. e Quaglino G. (Eds), *Dialoghi psicologici*. Unicopli.

Sohn D. (1996). Meta-analysis and science. *Theory & Psychology*, vol. (6): 229-247.

Trombetta C., Chirumbolo A. e Bruni D. (2001). Il contributo delle riviste italiane di psicologia alla diffusione delle tematiche psicoeducative. *Psicologia della formazione*, vol. (3): 383-402.

Walkerdine V., Lucey H. e Melody J. (2001). *Growing up girl*. New York University Press.

On Operationalizing Syntactic Complexity

Benedikt M. Szmrecsányi

University of Freiburg – Germany

benedikt.szmrecsanyi@uni-freiburg.de

Abstract

In the recent functional linguistic literature, the notion of syntactic complexity and similar concepts have received considerable attention. Yet, when trying to operationalize syntactic complexity as an independent variable in statistical research designs, it soon emerges that the notion is somewhat underdefined. I will report the results from an experiment comparing three measures of syntactic complexity — node counts, word counts, and a so-called 'Index of Syntactic Complexity' — with regard to their accuracy and applicability. While presupposing that node counts are cognitively the most 'real' measure, it turns out that the other two measures are near-perfect proxies of the former. I conclude that since node counts are in most cases unreasonably resource demanding to conduct, researchers can feel safe in using the measure that is most economically to conduct, word counts.

Keywords: stylometry, textual classification, text corpora and text encoding.

1. Introduction

What is meant by syntactic complexity (e.g. Ferreira, 1991; Givón, 1991), cognitive complexity (e.g. Mondorf, 2003; Givón, 1991; Rohdenburg, 1996), clause complexity (e.g. Kubon, 2001), linguistic complexity (e.g. Schleppegrell, 1992), structural complexity (e.g. Givón, 1991; Arnold *et al.*, 2000), or grammatical / syntactic weight (e.g. Wasow, 1997; Wasow and Arnold, 2003)[1] has received a good deal of scholarly attention, particularly — but not exclusively — by linguists working within functional frameworks. Yet, somewhat surprisingly, there is a dearth of precise definitions and convincing approaches to operationalize these concepts in a straightforward, objective, and non-intuitional way in empirical research designs. More often than not, the issue is avoided or side-stepped in that syntactically, cognitively, and/or structurally complex environments are sloppily 'defined' as contexts "that are for some reason more difficult, more complex, less entrenched, less frequent, less accessible or in any way cognitively more complex" (Mondorf, 2002: 252). Obviously, such definitions are not helpful when the task is to answer a variationist research question such as "How does the syntactic complexity of a slot's linguistic environment influence a speaker's decision to employ a pattern *x* rather than some other equivalent pattern *y*?". In this paper, therefore, I will report the results from an experiment comparing three measures of syntactic complexity — node counts, word counts, and a so-called 'Index of Syntactic Complexity' — with regard to their accuracy and applicability in empirical research designs.

[1] In this paper, all these terms will be considered synonymous and subsumed under 'syntactic complexity' since they all basically refer to syntactic structures which necessitate increased parsing and processing effort. Also note that due to space limitations, I will not be able to consider notions such as 'width' and 'depth' in relation to subordination.

2. Syntactic complexity of what?

A first question that has to be addressed is, what is the nature of the units of linguistic data to be compared? In most cases, a unit A of some sort will have to be compared to a unit B of some sort. To this end, units A and B must be roughly comparable in scope (for instance, it will hardly be operational to compare the syntactic complexity of a ten word utterance to the syntactic complexity of, say, a short story). In principle, then, it is possible to define scope in two different ways: (i) scope is defined in terms of pure length, duration, or size of a unit; (ii) scope is defined by appealing to notions independent of pure length, duration, or size. To illustrate, in approach (i) one could compare units of linguistic data which have the same number of words (for instance, syntactic complexity of one six word utterance is compared to syntactic complexity of another six word utterance); or which took speakers the same time to produce (for instance, a 10 second utterance is compared to another 10 second utterance); or that have the same size (for instance, one 100 byte chunk of text is compared to another 100 byte chunk of text). The main problem with such an approach is that appealing to length, duration, or size alone often makes impossible what is at stake here — using a structural measure of syntactic relationships in a unit of linguistic data. For illustration, consider (1) and (2):

(1) I wasn't there cause I had to fill out all this.

(2) I didn't do it, and the reason for this was that.

Both (1) and (2) have a length of 11 words, and are thus comparable in scope according to criterion (i). But while it is easy to see that (1) contains a main clause and a syntactically dependent adverbial clause of reason, (2) is a compound clause, as far as we can say with any reasonable degree of confidence. But the last word, *that*, might (e.g., *...the reason for this was that I was sick*) or might not (e.g., *the reason for that was that guy*) introduce a syntactically dependent complement clause. In other words, the measure of scope outlined here interferes with one's ability to assess syntactic complexity.

If, according to (ii), scope is defined by appealing to notions independent of pure length, one could compare units of linguistic data that have the same number of clauses, the same number of sentences, the same number of paragraphs, the same number of phrases, etc. To illustrate, if one were were to adopt the notion of 'sentence' as the basic unit of measure, one could compare (3) to (4), because both (3) and (4) constitute exactly one sentence:

(3) I wasn't there cause I had to fill out all this.

(4) I didn't do it, and the reason for that was that I was sick.

The major advantage of this approach, of course, is that notions such as 'clause', 'sentence', or 'paragraph' already have a requirement of meaningful, complete syntactic relationships built into them.

3. Previous approaches to assess syntactic complexity

Previous research has used a couple approaches to assess syntactic complexity (see, e.g., Wasow, 1997 on their advantages and disadvantages with regard to heavy noun phrase shift). They all basically take what was outlined above as approach (ii), i.e. they define the boundaries of what they measure by appealing to notions independent from length.

3.1. *Length as a proxy for syntactic complexity (number of words dominated)*

Arnold *et al.* (2000) and Hawkins (1990), for instance, measured syntactic weight of constituents as the difference in length (in words) between these constituents. Using length —

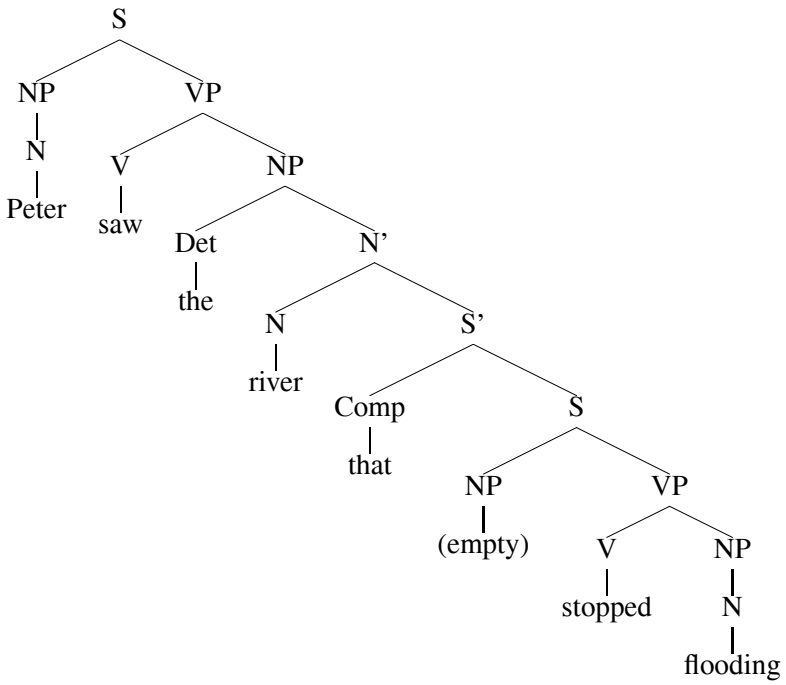
number of words, syllables, intonation units, etc. — as a proxy for syntactic complexity has the obvious advantage that this is a straightforward method which does not even necessarily involve manual coding. It is also probably the most time-honored proxy for complexity. Behagel (1909), in his seminal essay on what is today called 'the principle of end weight', already implicitly referred to it as a measure of weight. A drawback is that increased length does not necessarily always translate into increased syntactic complexity. For instance, concerning end weight, Chomsky (1975: 477) has remarked that "it is interesting to note that it is apparently not the length in words of the object that determines the naturalness of the transformation (i.e., heavy constituents shifted to final position, BS), but, rather, in some sense, its complexity."

If length is taken to mean the number of words dominated by a syntactic unit, then, this unit's syntactic complexity will be directly proportional to the number of words it contains. In what follows, the measure will be conceptualized by taking sentences as the basic unit and then determining sentence length (henceforth: SL). Thus, a sentence containing, for instance, 6 words will receive a syntactic complexity score of 6, while a sentence containing 9 words will receive a syntactic complexity score of 9.

3.2. *Counting nodes (number of nodes dominated)*

Johnson (1966), Ferreira (1991) and Rickford *et al.* (1995) counted nodes to determine syntactic complexity of phrases and sentences; Hawkins (1994) also suggested this method. In this approach, the more phrasal nodes a unit dominates, the more complex it is. Presupposing some notion of formal complexity, counting the number of nodes dominated is conceptually certainly the most direct and intuitively the most appropriate way to assess syntactic complexity. This is because the method reflects how the human parser is supposed to work. Thus, it is the structural measure which is psychologically most real. Note that syntactic structure — and complexity thereof — is not only a formal phenomenon, but also a a psychological one in that speakers must construct and produce material that conforms to the syntactic rules of their language. This is obviously more resource demanding for syntactically complex sentences than for syntactically simple sentences (for a discussion from a psycholinguistic perspective, see, for instance, Bock, 1982; Bock and Kroch, 1989; Ferreira, 1991; for a linguistic opinion, see Hawkins, 1990).

From a practical viewpoint, though, counting nodes almost always involves manual coding. This is the reason why the method is more popular in experimental research designs where only a very limited number of contexts has to be coded, in contrast to corpus-based designs where the number of contexts to code is potentially open-ended. If the method is adopted anyway, the idea is to measure the number of nodes in a unit of data using a parse tree.



To illustrate, the above tree diagram represents the sentence *Peter saw the river that stopped flooding*, which dominates 16 nodes (excluding the clause node S and the lexical items themselves) and therefore receives a syntactic complexity score of 16.

3.3. *An index of syntactic complexity (number of high complexity indicators dominated)*

A final approach to assessing syntactic complexity is conceivable. It would involve establishing what I would like to call an 'Index of Syntactic Complexity' (henceforth: ISC). For Beaman (1984: 45), "syntactic complexity in language is related to the number, type, and depth of embedding in a text. Syntactically simple authors [...] rely more heavily on coordinated structures [...] Syntactically complex authors [...] use longer sentences and more subordinate clauses." In her study, Beaman determines syntactic complexity of spoken and written discourse by simply comparing the percentage of subordinate clause structures in both discourse types. In a similar vein, Givón (1991) appears to consider embedded and/or subordinate structures — as opposed to conjoined structures — indicative of marked, and hence cognitively more complex, categories. Givón (1991: 347) cites a sizable body of psycholinguistic studies which demonstrate that subordinate clause structures are more complex to process than conjoined main clause structures.

Following Givón's and Beaman's emphasis on embeddedness, then, syntactic complexity of a given context could be established by counting linguistic tokens that can be considered telltale signs of increased grammatical subordinateness and embeddedness, such as (i) subordinating conjunctions (henceforth: SUB) (for instance, *because, since, as, when, that*, etc.), and (ii) WH-pronouns (henceforth: WH) (*who, whose, whom, which*). In addition, tokens to be included in the index should also include (iii) verb forms (henceforth: VF), both finite and non-finite, and (iv) noun phrases (henceforth: NP).

Because subordinators and WH-pronouns are the most straightforward indicators of increased embeddedness — and thus of high complexity —, these features should be weighted more heavily than verb forms and noun phrases. I would, then, like to suggest the following formula — which, admittedly, is somewhat tentative and ad-hoc — to establish (ISC):

$$\text{ISC}(u) = 2 \times n(u, SUB) + 2 \times n(u, WH) + n(u, VF) + n(u, NP)$$

Let *u* be the the unit of linguistic data under analysis, let ISC(*u*) be the ISC of the unit of linguistic data under analysis, and let n(*u*,SUB) be the number of occurrences of SUB in the unit of linguistic data under analysis, etc. According to this formula, ISC of a given unit of data is twice the number of occurrences of subordinating conjunctions and WH-pronouns plus the number of occurrences of verb forms and noun phrases in that unit. A structural measure such as ISC has the advantage that it is relatively easy to establish, especially when compared to counting nodes. If the data to be analyzed is part of speech (POS) tagged, it can be even done automatically by software. To illustrate ISC, consider (5):

> (5) The jury further said in term-end presentments that the City Executive Committee, which had over-all charge of the election, deserves the praise and thanks of the City of Atlanta for the manner in which the election was conducted.

This sentence would receive an ISC score of 21: it contains 1 subordinating conjunction (*that*), 2 WH-pronouns (*which, which*), 5 verb forms (*said, had, deserves, was, conducted*), and 10 NPs (*jury, term-end presentments, City Executive Committee, over-all charge, election, praise, thanks, City of Atlanta, manner, election*). Just for purposes of comparison, according to the SL measure, this sentence would receive a score of 38 (it has 38 words); according to the measure of how many nodes the sentence commands (henceforth: NODE), it would receive a score of 68 (it dominates 68 nodes, not including the clause node S and the lexical items themselves).

4. Comparing structural measures of syntactic complexity

4.1. *Method and data*

To systematically compare differences between the aforementioned approaches to measure syntactic complexity, each of the three measures was applied to the first 20 sentences of text A01 of the *Susanne Corpus* and the first 30 sentences of text T01 of the *Christine Corpus*, establishing syntactic complexity per sentence as analytical unit. The convenient advantage of the above corpora is that they are so-called treebanks, i.e. they are tagged for both POS and syntax. The Susanne corpus comprises a subset of the *Brown Corpus of American English* and is a corpus of written English. The Christine Corpus comprises a subset of the spoken-demographic section of the *British National Corpus* and is a corpus of spoken English (see http://www.cogs.susx.ac.uk/users/geoffs/SueDoc.html and http://www.cogs.susx.ac.uk/users/geoffs/ChrisDoc.html, respectively, for more information on these corpora). Both corpora are syntactically annotated using a phrase structure approach to English grammar.

4.2. *Results*

Numerical structural complexity scores returned by different structural measures are not cardinally comparable. They are comparable in an ordinal fashion only, i.e. how they rank different sentences with regard to syntactic complexity. This is what is indicated in brackets in table 1. When a sentence occupies rank 1, this means that there is no sentence that is syntactically more complex in the dataset, while the lowest rank indicates that there is no sentence which is syntactically less complex in the dataset. Note that because two or more sentences may receive identical syntactic complexity scores, more than one sentence may occupy a given rank. This is why the absolute number of ranks varies across different structural measures. Figure 1, then, visualizes the results of the three structural measures by indicating relative ranks to enhance

	Christine corpus			*Susanne corpus*		
sentence #	SL	NODE	ISC	SL	NODE	ISC
1	7 (8)	17 (7)	3 (4)	21 (10)	39 (14)	12 (7)
2	6 (9)	11 (11)	1 (6)	38 (3)	68 (3)	21 (1)
3	6 (9)	11 (11)	1 (6)	33 (4)	56 (5)	14 (5)
4	17 (2)	34 (2)	5 (2)	29 (6)	49 (7)	13 (6)
5	2 (12)	2 (17)	0 (7)	21 (10)	41 (12)	9 (10)
6	3 (11)	7 (14)	1 (6)	21 (10)	44 (9)	11 (8)
7	2 (12)	2 (17)	0 (7)	39 (2)	76 (2)	16 (3)
8	5 (10)	11 (11)	1 (6)	22 (9)	40 (13)	11 (8)
9	7 (8)	12 (10)	2 (5)	20 (12)	36 (16)	7 (12)
10	16 (3)	33 (3)	9 (1)	11 (15)	22 (19)	8 (11)
11	14 (4)	27 (5)	4 (3)	13 814)	26 (18)	4 (13)
12	2 (12)	3 (16)	0 (7)	25 (7)	42 (11)	14 (5)
13	19 (1)	35 (1)	9 (1)	23 (8)	43 (10)	8 (11)
14	9 (7)	15 (8)	4 (3)	52 (1)	90 (1)	18 (2)
15	6 (9)	12 (10)	2 (5)	20 (11)	36 (16)	9 (10)
16	6 (9)	9 (13)	1 (6)	20 (11)	37 (15)	11 (8)
17	6 (9)	12 (10)	2 (5)	14 (13)	28 (17)	7 (12)
18	7 (8)	14 (9)	3 (4)	32 (5)	61 (4)	15 (4)
19	1 (13)	1 (18)	0 (7)	29 (6)	48 (8)	10 (9)
20	1 (13)	1 (18)	0 (7)	29 (6)	54 (6)	11 (8)
21	6 (9)	10 (12)	3 (4)			
22	1 (13)	1 (18)	0 (7)			
23	3 (11)	7 (14)	2 (5)			
24	2 (12)	5 (15)	1 (6)			
25	11 (5)	21 (6)	3 (4)			
26	10 (6)	17 (7)	2 (5)			
27	14 (4)	29 (4)	3 (4)			
28	2 (12)	5 (15)	1 (6)			
29	6 (9)	11 (11)	2 (5)			
30	5 (10)	9 (13)	1 (6)			

Table 1. Syntactic Complexity scores (ranks in brackets) according to different structural measures in the Susanne Corpus (written English) and the Christine Corpus (spoken English)

comparability. This means that ranks returned by the NODE measure were taken as baseline and the ranks returned by the other two measures were then plotted with regard to their relative relation to the NODE measure ranking. Sentences on the X axis were ordered according to the baseline ranking suggested by the NODE measure.

A cursory glance at these tables and especially the graphs indicates that the three structural measures rank the data in an indeed surprisingly conform fashion. Assuming — as it was done here — that NODE is conceptually and psychologically the 'most real' structural measure, it can be seen that neither SL nor ISC produce substantially significant outliers in either dataset. This visual impression is confirmed when Spearman's correlation coefficient (*Spearman's* ρ) is computed. This statistical procedure provides a means to assess the linear relationship between two variables. The resulting coefficient can range between -1 (which would indicate a perfect negative linear relationship) and +1 (which would indicate a perfect positive linear relationship). Actual correlation coefficients can be seen from table 2.

Overall, SL and ISC are better proxies for NODE in spoken data than in written data. Also, SL slightly outperforms ISC in both datasets. It should be pointed out though that with Spearman's

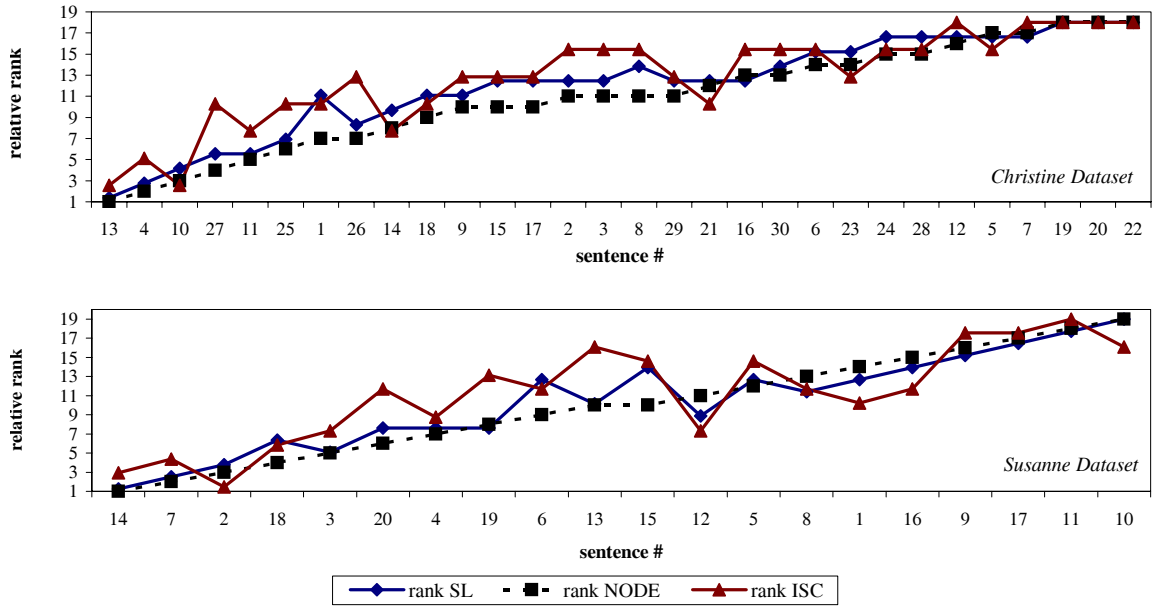


Figure 1. Relative Syntactic Complexity ranks according to different structural measures in the Susanne corpus (written English) and Christine Corpus (spoken English)

	SL VS. NODE	ISC VS. NODE
Susanne corpus (written English)	.976**	.836**
Christine corpus (spoken English)	.989**	.916**

** significant at the .01 level

Table 2. Correlation coefficients for rankings suggested by different structural measures

ρ in the .8/.9 range, both measures are highly correlated to NODE. Both SL and ISC, therefore, are near-perfect proxies of NODE.

5. Summary and conclusion

In conclusion, I have suggested that measuring sentence length (SL) and, albeit to a slightly lesser degree, computing an index of syntactic complexity (ISC) do an excellent job in approximating a node count (NODE), the structural measure of syntactic complexity which is probably the most 'real' one cognitively. In essence, thus, the three structural measures really gauge the same thing. Interestingly, Wasow (1997) has found to be much the same true for measuring weight effects ("It is very hard to distinguish among various structural weight measures as predictors of weight effects. Counting words, nodes, or phrasal nodes all work well", Wasow, 1997: 102). Importantly, though, the three structural measures presented here differ markedly in what data they require and in the expenditure of manual coding they necessitate. SL just requires counting words, which even any standard word processor is able to perform without human help. ISC requires manual coding, unless the data source is POS tagged. NODE also requires manual coding, unless the data source is syntax tagged. More often than not, the researcher is confronted with data that are not tagged at all. What I hope to have provided evidence for in this paper is that the most economic method — determining length in words — to assess

syntactic complexity is by all means one that is nearly as accurate as the more sophisticated and cognitively, conceptually, or even psychologically 'more real' methods.

References

Arnold J., Wasow T., Losongco A. and Ginstrom R. (2000). Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. *Language*, vol. (17/1): 28-55.

Beaman K. (1984). Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse. In Tannen D. and Freedle R. (Eds), *Coherence in Spoken and Written Discourse*: 45-80.

Behagel O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, vol. (25/110).

Bock K. (1982). Towards a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, vol. (89): 1-47.

Bock K. and Kroch A. (1989). The isolability of syntactic processing. In Carlson G. and Tanenhaus M. (Eds), *Linguistic structure in language processing*.

Chomsky N. (1975). *The logical structure of linguistic theory*. Chicago University Press.

Ferreira F. (1991). Effects of Length and Syntactic Complexity on Initiation Times for Prepared Utterances. *Journal of Memory and Language*, vol. (30/2): 2110-2233.

Givón T. (1991). Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language*, vol. (15/2): 335-370.

Hawkins J. (1990). A Parsing Theory of Word Order Universals. *Linguistic Inquiry*, vol. (21/2): 223-261.

Hawkins J. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press.

Johnson N. (1966). On the relationship between sentence structure and the latency in generating the sentence. *Journal of Verbal Learning and Verbal Behavior*, vol. (5): 375-380.

Kubon V. (2001). A Method for Analyzing Clause Complexity. *Prague Bulletin of Mathematical Linguistics*, vol. (75): 5-28

Mondorf B. (2003). Support for More-Support. In Rohdenburg G. and Mondorf B. (Eds), *Determinants of Grammatical Variation in English*: 251-304.

Rickford J., Denton M., Wasow T. and Espinoza J. (1995). Syntactic Variation and Change in Progress: Loss of the Verbal Coda in Topic-Restricting *As Far As* Constructions. *Language*, vol. (71/1): 102-131.

Rohdenburg G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics*, vol. (7): 149-182.

Schleppegrell M. (1992). Subordination and Linguistic Complexity. *Discourse Processes: A Multidisciplinary Journal*, vol. (15/1): 117-131.

Wasow T. (1997). Remarks on grammatical weight. *Language Variation and Change*, vol. (9): 81-105.

Wasow T. and Arnold J. (2003). Post-verbal constituent ordering in English. In Rohdenburg G. and Mondorf B. (Eds), *Determinants of Grammatical Variation in English*: 119-154.

Cohésion informative dans le discours scientifique

Cristina Alice Toma

Faculté des Lettres – CH-1211 Genève – Suisse
toma1@etu.unige.ch

Abstract

The didactic discourse and the vulgarization discourse are two varieties of discourse. Knowing and understanding the structure of the scientific discourse are essential for a better transmission of knowledge. The principal dimension of the scientific discourse involved in the transmission of knowledge is the informative one. The purpose of this work is to describe some aspects of the informational organization and of the topical organization of the scientific discourse. We pay a special attention to the scientific terms, since the investigation of their role in the topical organization of the scientific discourse enables us to see the relationship between terminology and the discourse analysis.

Résumé

Le discours didactique et de vulgarisation sont des variantes du discours scientifique. Connaître et comprendre le fonctionnement de ce dernier, c'est améliorer la transmission des connaissances. La dimension principale du discours scientifique impliquée dans le processus de transmission de connaissances est celle informative. Nous nous proposons de relever quelques aspects de l'organisation informationnelle et topicale du discours scientifique qui nous aide à la mise en contexte des concepts – termes scientifiques. La « contextualisation » des termes est nécessaire pour une meilleure maîtrise de ces derniers[1].

Mots-clés : discours scientifique, terme, analyse du discours, organisation informationnelle, représentation conceptuelle, organisation topicale.

1. Introduction

Pourquoi l'intérêt pour le discours scientifique ?

Premièrement, parce que le discours scientifique a une importance très grande dans le développement de la science aujourd'hui. Selon Michael Halliday et James R. Martin, la science est « une technologie du discours » (1993 : 10). Voilà pourquoi une étude linguistique du discours (pour la définition v. Roulet, 1986 et 2001) scientifique ne manquerait pas d'intérêt pour les scientifiques en général.

Deuxièmement, parce que de nombreuses recherches portant sur le langage scientifique s'arrêtent à l'étude des termes qui, même s'ils constituent une caractéristique importante du langage scientifique, n'en sont pas la seule. C'est pourquoi nous considérons qu'une analyse du discours scientifique pourrait, d'une part, venir à l'encontre de la terminologie pour lui offrir un cadre élargi, d'autre part, fournir ses résultats à l'intelligence artificielle, au traitement automatique des langues.

[1] Ce travail fait partie de notre thèse ayant comme directeur le prof. Eddy Roulet, que nous remercions beaucoup pour ses remarques et ses suggestions permanentes.

Pourquoi l'intérêt pour la dimension informationnelle du discours scientifique ?

Premièrement, parce que la fonction informative occupe un rôle central parmi les fonctions du discours scientifique. Michael Halliday et James R. Martin (1993) soutiennent que les langages et les discours des sciences ont vraiment des propriétés caractéristiques qui ont évolué pour pouvoir prendre en charge différentes tâches du travail cognitif et sémiotique que le langage du « sens commun » de chaque jour ne peut pas remplir ; par exemple, la représentation de la technicité et de l'abstraction.

Deuxièmement, parce que les termes scientifiques sont les véhicules principaux de l'information dans le discours scientifique.

Nous nous proposons d'analyser la gestion de l'information dans le discours scientifique et l'insertion des concepts – termes (pour leur définition v. Lerat, 1995 ; Cabré, 1998 ; Gaudin, 1998 et 2003). Pour ce faire nous prenons un exemple qui soit à la porté des linguistes et spécialistes à la fois. Il s'agit du discours scientifique de vulgarisation. Il faut rappeler ici que le discours scientifique n'est pas une espèce monolithique. Au contraire, on distingue, à partir de différents facteurs (cf. le cadre interactionnel et actionnel de l'analyse du discours genevoise, Roulet, 2001), au moins trois niveaux du discours scientifique : le discours scientifique proprement dit ou de recherche, le discours scientifique didactique et le discours scientifique de vulgarisation.

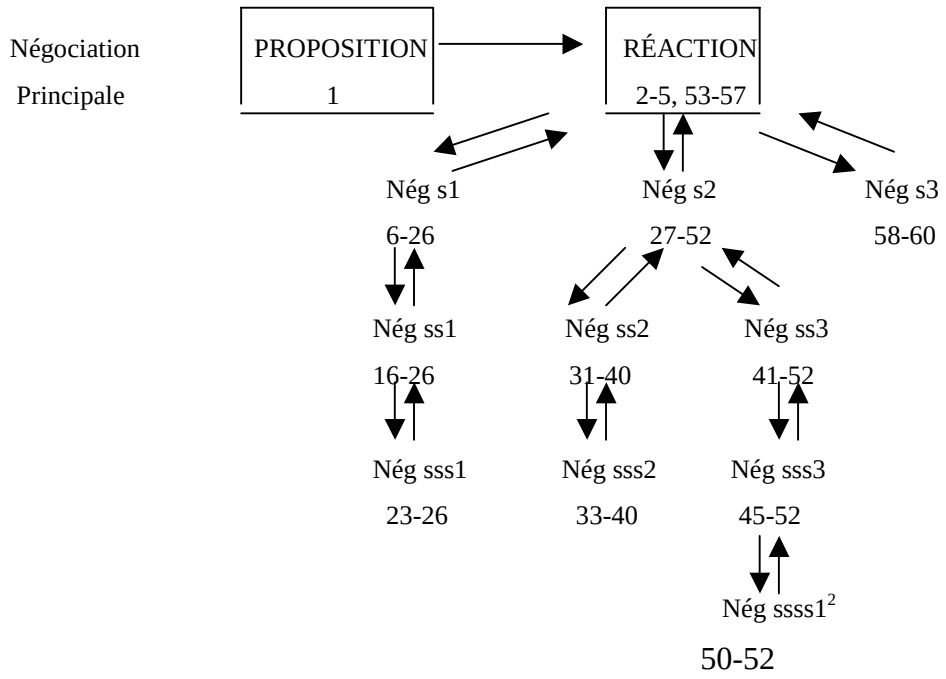
Nous relèverons quelques aspects de l'organisation informationnelle et topicale du discours scientifique, en montrant la continuité (progression à topique constant ou progression linéaire) ou la discontinuité (enchaînement à distance) dans la construction du savoir. Nous montrerons aussi comment les réseaux de concepts se construisent dans le discours scientifique et comment les termes s'« emboîtent » les uns les autres, fait qui constitue d'ailleurs une de leurs propriétés. Pour ce faire nous utilisons la méthode offerte par le modèle genevois d'analyse du discours en procédant, tour à tour, à l'étude de la structure hiérarchique, de l'organisation informationnelle, de la structure conceptuelle et de l'organisation topicale du discours scientifique.

2. La structure hiérarchique

Le modèle genevois d'analyse du discours est un *modèle modulaire*. Les trois dimensions du discours (*linguistique, textuelle et situationnelle*) sont étudiées par l'intermédiaire de cinq modules (*lexical, syntaxique, hiérarchique, référentiel, interactionnel*) parmi lesquels la structure hiérarchique occupe une place centrale. Pour rendre compte des carctéristiques complexe du discours, on utilise le mécanisme du *couplage* des modules et/ ou des formes d'organisation simple qui engendre les *formes d'organisation élémentaires* (par exemple, l'organisation informationnelle), respectivement, *les formes d'organisation complexes* (par exemple, l'organisation topicale) (v. Roulet *et al.*, 2001).

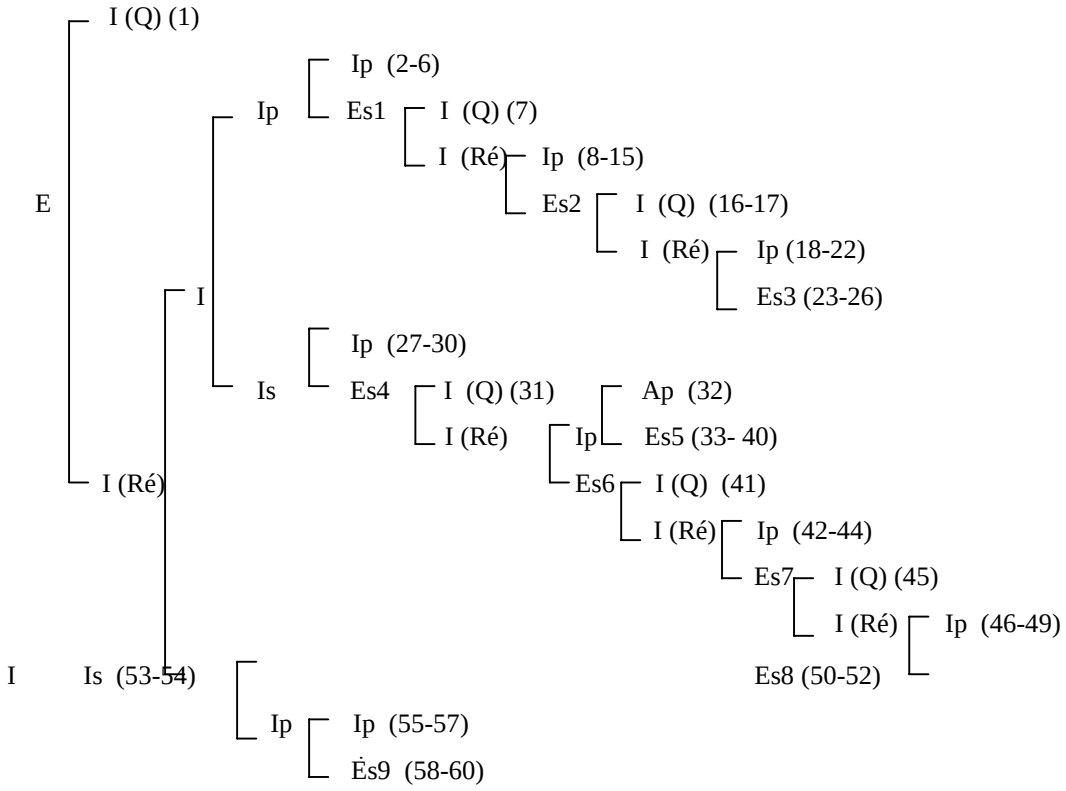
Pour décrire la structure hiérarchique d'un discours il faut comprendre le processus de négociation sous-jacent : « Puisque celle-ci [la structure hiérarchique] est la face émergente du processus de négociation sous-jacent a l'interaction, il faut commencer par décrire le parcours de cette négociation. » (Roulet, 2000 : 211).

Si l'on applique l'organigramme général de la négociation à ce fragment d'interaction, on obtient la structure suivante :



La négociation principale de l'échange porte sur Q*u'est-ce que les mathématiques pour vous [A. L.]*. Elle comprend plusieurs négociations enchâssées subordonnées, de différents degrés.

À ces processus de négociation correspond la macro-structure suivante :



[2] On a convenu de noter dans chaque négociation les négociations que celle-ci comprend. On a noté : Nég = négociation, s = secondaire ; le nombre de répétitions de s montre le degré de subordination.

Le fragment analysé représente une suite de 9 échanges subordonnés de différents degrés qui forment un grand échange. La structure explicite de chaque échange est binaire, du type question – réponse, mais chaque fois la ratification est implicite[3]. On ne peut pas passer à une autre négociation sans l'accord qui assure la clôture de la négociation précédente.

L'asymétrie des deux interventions du grand échange est frappante. L'intervention de question est formée d'un seul acte, tandis que l'intervention de réponse est formée de 59 actes et elle a une structure monologique[4], structure permise par le principe de récursivité. Cette asymétrie correspond au but du fragment : construire une définition des mathématiques du point de vue du mathématicien A. L. La construction de la définition devient finalement une co-construction, fait qui est la raison pour laquelle à quelques points (l'acte [16] et l'acte[22]) nous avons choisi l'interprétation des échanges Es2 et Es3 comme étant subordonnés et non pas au même niveau que la question principale. Ils constituent des morceaux de ce processus de co-construction qui traverse tout l'échange, tout ce fragment[5].

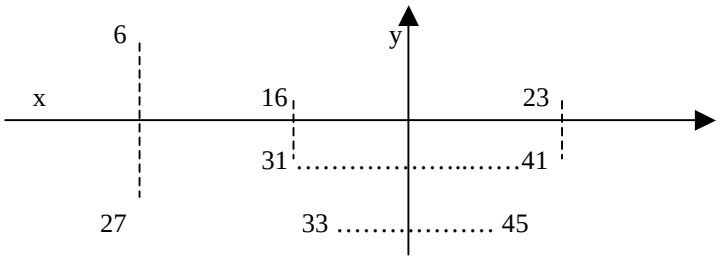
La macro-structure hiérarchique s'articule de façon symétrique. Les deux membres de la symétrie sont des interventions qui occupent une position de même degré dans la structure hiérarchique. La ressemblance de niveau de la position de profondeur dans la structure hiérarchique des interventions symétriques s'associe au type de composante de l'échange, principale ou secondaire. Par exemple, l'intervention réactive comprend deux interventions coordonnées principales (à partir de l'acte 2, respectivement, 5). D'une manière semblable se relient les interventions qui commencent à l'acte 6 et, respectivement, à l'acte 27.

À partir de ce niveau, la symétrie se manifeste à l'intérieur de deux interventions et en même temps entre les deux interventions. On pourrait dire qu'il s'agit d'une double symétrie. On peut représenter cette symétrie à l'aide du schéma suivant :

[3] Le discours comporte trois unités : *l'acte* (A) – la plus petite unité, *l'intervention* (I) et *l'échange* (E) – la plus grande unité qui comprend trois interventions : *la question* (Q), *la réponse* (Ré) et *la ratification* (Ra). Chaque unité (sauf l'échange qui est seulement subordonné) peut être principale (p) ou subordonnée (s). La structure hiérarchique est construite sur le principe de *récursivité* qui permet l'intégration d'une unité plus grande dans une unité plus petite.

[4] L'intervention est définie dans l'École genevoise comme l'unité attribuée à un locuteur unique, mais qui peut être composée elle-même d'autres interventions composant un échange. Ce fait est explicité par le principe de récursivité et est à l'origine de la distinction entre discours *dialogal* (plusieurs locuteurs) et le discours *dialogique* (structure d'échange composé de plusieurs interventions) d'une part et entre discours *monologal* (un seul locuteur) et discours *monologique* (structure d'intervention qui peut impliquer plusieurs locuteurs) d'autre part. (cf. Roulet *et al.*, 1985 : 72)

[5] La question est clairement et complètement formulée, même si elle ne comprend qu'un seul acte [1], fait qui lui assure la complétude monologique et qui empêche le prolongement vertical de l'échange. La réponse se construit au fur et à mesure à travers un développement horizontal de l'échange. A part la longueur, on pourrait préciser encore deux particularités de la réponse. Premièrement, la construction à distance des deux interventions coordonnées (2-53) – 51 actes séparent la première paraphrase « définitionnelle » *c'est très compliqué… Les mathématiques… c'est une science hors de la science* d'une deuxième paraphrase « définitionnelle ». *Les mathématiques, si on les reprend à leur origine sont nées en même temps que la philosophie* – montre l'unité de l'échange autour de la définition des mathématiques. Deuxièmement, l'hésitation dans l'intervention réactive de l'interlocuteur qui, dans un premier temps (acte 2), donne une réponse négative *c'est très compliqué* paraphrasable par « je ne sais pas vous répondre » , mais, vu l'engagement conscient pour la co-construction du dialogue, l'interlocuteur continue, dans un deuxième temps, par l'acte 3 et les actes suivants avec une réponse positive.



Pour simplifier j'ai gardé dans le schéma seulement l'acte principal.

X = axe de symétrie entre les interventions ;

Y = axe de symétrie à l'intérieur de l'intervention.

Cette structure symétrique vient à l'appui de la co-construction et, en même temps, elle constitue l'émergence de ce type de construction. La symétrie va avec les relations qui s'établissent entre les composantes de l'échange et, en plus, elle assure la possibilité d'une anticipation de la question du locuteur de la part de l'interlocuteur, mais aussi de la part de l'analyste[6].

3. Organisation informationnelle

L'organisation informationnelle rend compte de la continuité informative du discours en couplant des informations d'ordre hiérarchique et linguistique ou référentiel[7]. Elle doit identifier les topiques, les objets de discours (propos) et relever les différents types de progression de l'information[8].

Voilà l'analyse de l'organisation informationnelle de notre fragment[9] :

Organisation informationnelle

1. Qu'est-ce que les mathématiques **pour vous** ?
2. *L* — … **c**'[qu'est-ce que les mathématiques pour vous] est très compliqué.
3. *L* (très compliqué) Les mathématiques…
4. *L* **c**'[les mathématiques] est une science hors de la science,
5. *TC* **c**'[les mathématiques] est une science qui participe à l'art,
6. *TC* donc **qui** [les mathématiques] peut donner des satisfactions aussi bien esthétiques qu'intellectuelles…
7. *L* (donner des satisfactions)— esthétiques…
8. *L* (esthétiques) — Il y a une esthétique des mathématiques.

[6] La possibilité de l'anticipation de la question assure la cohérence et l'enrichissement progressif des informations introduites dans le discours (voir 6 et 8). Elle est facilitée par le fait que la question est extraite de la réaction précédente. Par exemple, la question formulée dans l'acte *31. Il y avait un certain plaisir, autrement dit, à se cogner contre quelque chose* est tirée de l'acte *30. C'était que je me cognais durement ;* l'information de l'acte *41. Se battre avec son esprit, autrement dit* est tirée de l'acte *36. Vous vous battez avec votre esprit fonctionnant dans les réalités.*

[7] « L'organisation informationnelle résulte du couplage entre des informations hiérarchiques et des informations lexicales ou syntaxiques, si l'acte comporte des traces de point d'ancrage du topique (pronoms, expressions définit, etc.), ou, en l'absence de telles traces, du couplage entre des informations hiérarchiques et des informations référentielles. » (Roulet, 1999 : 58)

[8] « Chaque acte introduit une information dite activée, l'objet de discours, et l'introduction de cette information implique au moins un point d'ancrage en mémoire discursive, le topique, sous la forme d'une information semi-active, qui peut être verbaliser ou non .» (Roulet , 2000 : 225)

[9] Nous mettons en gras *la trace topicale*, entre parenthèses - *le topique implicité* et entre crochets – *le topic marqué par une trace topicale*. Les numéros en tête de ligne comptent les actes, tandis que les lettres montrent le type de *progression informationnelle* (L = *linéaire, avec topic constant* ; = *à distance*).

9. *L*(esthétique des mathématiques) Je dis parfois que les mathématiques sont l'art le plus abstrait qui soit…
10. *L* (l'art le plus abstrait) on sait bien que les mathématiques emploient des termes,
11. *L* (les mathématiques) font de belles démonstrations,
12. *TC* (les mathématiques) ont une forme d'élégance
13. *L+D* et **ça** [10-12] correspond à la culture d'une certaine sensibilité intellectuelle qui n'est pas tellement différente de la sensibilité,
14. *L* (sensibilité) mettons,
15. *TC* (sensibilité) musicale. […]
16. *L(*musicien**)** — Vous ,
17. *L* **vous** étiez musicien aussi ?
18. *L* (musicien) — Non,
19. *TC* (musicien) très peu ;
20. *D* […] **J'**[ANDRÉ LICHNEROWICZ] avais une très mauvaise santé…
21. *L* et **ce qui** [une très mauvaise santé] était très agréable à certains points de vue…
22. *TC* (une très mauvaise santé) m'a permis de prendre l'habitude de travailler et de réfléchir seul, […]
23. *D* — Et vous voyez **une relation entre les mathématiques et cette période de votre enfance**, ?
24. *L* (une relation entre les mathématiques) cette façon de travailler seul…
25. *L* (22-23) — Peut-être…
26. *TC* (22-23) non ;
27. *D* (honnêteté intellectuelle)la vérité
28. *L* **c**[la vérité]'est que les mathématiques m'ont toujours apporté une joie d'honnêteté intellectuelle. […]
29. *L* **Ce que**[c'était que je me cognais durement] je trouvais extrêmement agréable en mathématiques,
30. *L***C**[je trouvais extrêmement agréable en mathématiques]'était que je me cognais durement ; […]
31. *L* — Il y avait un certain plaisir, autrement dit, **à se cogner contre quelque chose**.
32. *L* — **C**'[de se cogner] était un très grand plaisir **de se cogner**.
33. *L* — (un très grand plaisir) — Pourquoi ?
34. *L* — (pourquoi) Pourquoi…
35. *D* (un très grand plaisir) parce que vous ne vous battez pas avec des fantômes…
36. *L* **Vous vous battez** avec votre esprit fonctionnant dans les réalités
37. *L* (Vous vous battez avec votre esprit fonctionnant dans les réalités) et quand les choses ne vont pas,
38. *L* vous vous **en** [les choses ne vont pas] apercevez durement.
39. *L* **Ce qui** [37] est un motif de sécurité
40. *TC* (37)et non pas d'insécurité.
41. *L+D* (un motif de sécurité et non pas d'insécurité) — Se battre avec son esprit, autrement dit.
42. *L* (se battre avec son esprit) — Se battre avec son esprit,
43. *TC* (Se battre avec son esprit) dans la mesure où votre esprit est l'esprit de tout le monde.
44. *L* **Le fait que votre esprit ait vocation universelle** est probablement très sécurisant.
45. *L* (très sécurisant) — Pourquoi ?
46. *L* (pourquoi)— Pourquoi ? … (silence)…
47. *D* (très sécurisant) Parce que vous n'êtes pas victime de mythes ou de fantasmes…
48. *L* **que**[de mythes ou de fantasmes] vous savez mal apprécier.
49. *D* Ce sont les mathématiques qui **ont d'ailleurs donné à l'humanité la notion même de probité intellectuelle**.
50. *TC* — Qu'est-ce que vous mettez sous **ces mots de probité intellectuelle** ?
51. *TC* **Ils** [ces mots de probité intellectuelle] paraissent importants pour vous.
52. *D* (vous mettez) — Ne jamais être dupe de soi-même ou des autres.
…(silence)…
53. *D* (1) — Les mathématiques,

54. *L* si on **les** [les mathématiques] reprend à leur origine
55. *TC* (les mathématiques) sont nées en même temps que la philosophie,
56. *L* s'**en** [philosophie] sont différenciées,
57. *D* car **elles** [les mathématiques] ont cherché à établir… un type de discours cohérent, contraignant pour l'autre et sans bruit de fond, sans quiproquos ni malentendus.
58. *L* (un type de discours) — Vous dites, contraignant pour l'autre,
59. *L* (contraignant pour l'autre) comment ?
60. *D* — Parce qu'**il** [type de discours] est capable, par sa forme, d'interdire le refus de son contenu.[…].

On commence par l'identification des topiques et des propos[10] (par défaut ou par le reste de la soustraction des topiques) et on continue avec les types de progression d'information.

Les topiques constituent l'information la plus accessible semi-activée[11], des points d'ancrage immédiat[12] qui ont pour source soit l'avant-texte, soit l'environnement cognitif immédiat, soit une inférence tirée de ce deux première source. Le topique est verbalisé par une trace topicale (anaphore ou expression définie) ou reste implicité (dans notre analyse on indique le topique implicité en tête d'acte, entre parenthèses). Le test qu'on peut utiliser pour identifier le topique sont : le détachement à gauche, le clivage ou le paraphrase du type « à propos de… »[13].

Par exemple, le test de détachement à gauche aide à l'identification du topique du premier acte :

1. Qu'est-ce que les mathématiques pour vous ?
1. Pour vous qu'est-ce que les mathématiques ?

En ce qui concerne le topique du deuxième acte, on peut hésiter entre *les mathématiques* et la question *Qu'est-ce que les mathématiques pour vous ?*

Cette fois-ci le choix (la deuxième solution) et motivé par les connaissances encyclopédiques : les mathématiques ne sont pas compliquées pour un mathématicien, mais définir ce qu'elle représente pour un mathématicien peut être compliqué. On utilise l'environnement cognitif immédiat pour identifier le topique de l'acte *33 : un très grand plaisir*. On fait appel à la réponse pour expliquer le topique de la question étant donné qu'il existe toujours un enchaînement informationnel entre une question et sa réponse :

35. se cogner parce que vous ne vous battez pas avec des fantômes.*
35. un très grand plaisir parce que vous ne vous battez pas avec des fantômes.

On constate qu'il y a un nombre important de traces tropicales (presque la moitié des topiques), ce qui prouve le caractère hybride de notre texte, situé entre un dialogue où le traces topicales manquent et un texte scientifique où les traces topicales sont nombreuses.

Quant aux topiques implicités, ils constituent une pré-construction de la structure conceptuelle

[10] La notion de *propos* est équivalente à la notion d' *objet de discours*. Elle a été introduite dans les dernières recherches de l'École genevoise (2001).

[11] On part de la distinction établit par Chafe (1994 : 53-56) entre information inactive, semi-active et active (ou, mieux, activée) (cf. Roulet, 2000 : 224).

[12] Dans la mesure où le locuteur constitue un point d'ancrage disponible tout au long du dialogue on considère qu'il constitue un point d'ancrage d'arrière-fond, par opposition au topique, qui peut être décrit comme un « point d'ancrage immédiat » (cf. Roulet *et al.*, 2001 : 253).

[13] Un cas particulier de la source d'un topique est celui de l'acte 29. Ce topique prend comme source l'objet de discours de l'acte suivant 30. Il s'agit d'une cataphore.

29. Ce que (que je me cogner durement) j'ai trouvé extrêmement agréable en mathématiques.

du texte (*satisfactions esthétiques, art, sensibilité, musicien, plaisir, se battre avec son esprit, sécurisant, mathématiques, type de discours*). Une possible explication serait le fait que le thème de l'entretien est si clair que son développement va de soi.

À partir du critère de l'origine du topique, il est possible de déterminer les principaux types de progressions informationnelles : progression linéaire, si le topique est tiré du propos qui précède ; progression à topique constant, si le topique est tiré du topique de l'acte précédent ; l'enchaînement à distance, si le topique est tiré d'un autre acte que l'acte précédent[14].

L'enchaînement à distance assure le retour en arrière pour reprendre les propos ou les topiques principaux du dialogue qui ne sont plus directement accessible dans la succession immédiate de l'information (*20-21, 26-27, 34-35, 46-47, 51-52, 52-53, 59-60).* Si les enchaînements *34-35, 46-47* apparaissent comme un effet d'une reprise de la question par l'interlocuteur, ce qui détermine le saut par dessus un acte du topique, les enchaînement à distance *51-52* et *59-60* sont le résultat de la formulation de la question en deux temps. Les enchaînement à distance *26-27* et *52-53* reviennent sur les propos *satisfaction intellectuelle* et, respectivement, *les mathématiques*.

On constate un va-et-vient entre les propos et les topiques. Presque tout propos devient à un moment donné le topique d'un autre acte. Le fait s'explique par le besoin de précision que le locuteur et l'interlocuteur cherchent dans la construction de leur discours[15].

L'organisation informationnelle part de l'hypothèse qu'il existe une continuité absolue de l'information dans le texte. Mais on constate pourtant qu'il y a à travers le discours des noyaux informationnels indépendants dont l'existence suffit à eux-mêmes. Il s'agit des paires de deux actes où le topique d'un acte est tiré du propos d'un autre et à l'envers, par phénomènes opposés de cataphore et, respectivement, anaphore. Par exemple :

27. (honnêteté intellectuelle) la vérité

28. c'(la vérité) est que les mathématiques m'ont toujours apporté une joie d' honnêteté intellectuelle[16]

29. ***ce que*** *(c'était que je me cogner durement) je trouvais extrêmement agréable en mathématiques*

30. ***c'*** *(je trouvais extrêmement agréable en mathématiques) était que je me cogner durement*

C'est l'effet circulaire de l'information qui se clore sur elle-même.

---

[14] Le type dominant de progression est celui de la progression linéaire : *1-2, 2-3, 3-4, 6-7, 7-8, 8-9, 9-10, 10-11, 13-14, 15-16, 20-21, 24-25, 26-27, 31-32, 32-33, 34-35, 35-36, 36-37, 37-38,38-39, 41-42, 43-44, 44-45,45-46, 53-54, 55-56, 57-58, 58-59.*

On reste dans le domaine de la progression linéaire pour les actes *15-16, 30-31, 43-44.* Ce qui est différent de l'objet de discours de l'acte précédent (objet de discours source) , ce n'est pas le topique, mais son actualisation linguistique. Par exemple, *musicale – musicien ; je me cognais durement – se cogner contre quelque chose.*

La progression linéaire s'associe à l'enchaînement à distance lorsque le topique a comme source deux ou trois actes qui précèdent (*12-13, 40-41).*

La progression à topique constant apparaît entre les actes : *4-5, 5-6, 11-12, 14-15, 15-16, 18-19, 21-22, 23-24, 25-26, 30-31, 42-43, 49-50, 50-51, 54-55.* On retrouve la progression à topique constant au passage d'une réponse à la question suivante (*30-31, 49-50*), ce qui vient à l'appui de l'hypothèse de co-construction consciente et active du discours par les deux participants à l'entretien.

[15] Une remarque supplémentaire s'impose pour l'acte *20. J'avais une très mauvaise santé* et l'enchaînement à distance que nous y avons signalé. *Je* peut être la trace d'un point d'ancrage d'arrière-fond ou la trace d'un point d'ancrage immédiat. Nous avons opté pour la deuxième solution parce que ce *je* ne désigne seulement le locuteur de discours, mais ici il désigne premièrement le mathématicien qui définit les mathématiques et qui constitue le topique du premier acte.

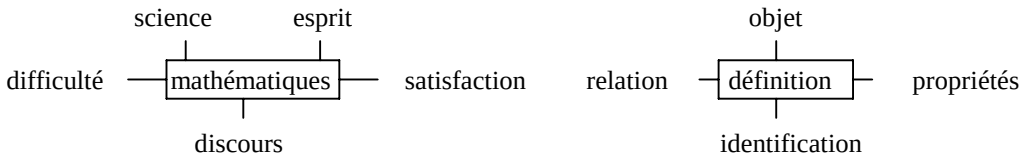
[16] Pour ces deux actes on a accepté aussi l'interprétation de continuité informationnelle (enchaînement à distance).

4. Dimension référentielle. Représentation et structure conceptuelle

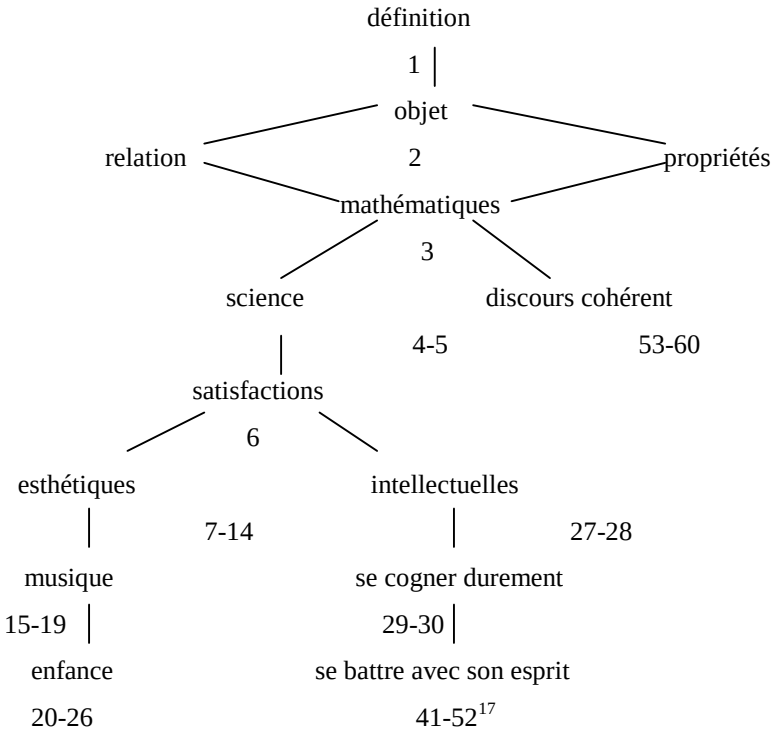
Le discours se trouve au croisement du monde dans lequel il s'inscrit et du monde dont il parle. L'analyse du discours doit rendre compte du cadre actionnel du discours, des représentations et des structures praxéologiques et conceptuelles des actions et des objets (cf. Roulet, 1999 : 52).

Dans le discours scientifique, ce qui compte ce sont les concepts. C'est pourquoi nous ne proposons ici que l'analyse des objets dont le discours parle. Les concepts principaux de notre discours sont *mathématiques* et *définition. Mathématiques* représente l'objet- patient verbalisé dans le discours. *Définition* représente l'objet- opérateur, non-verbalisé dans le discours, mais implicité par des expressions comme *qu'est-ce que* ou *c'est*.

On peut schématiser les représentations conceptuelles prototypiques des *mathématiques* et de la *définition* ainsi :



Dans l'échange analysé ici ces représentations prototypiques sont combinées et exploitées d'une manière particulière, qu'on peut schématiser dans la structure conceptuelle suivante :



5. L'organisation topicale

L'organisation topicale, forme d'organisation complexe, traite les liens conceptuels entre les objets de discours. Le couplage de la structure hiérarchique et de l'organisation information-

[17] On indique ici aussi le parcours informationnel en précisant les actes qui activent les concepts mentionnés.

nelle rend compte de l'enchaînement à distance et de la hiérarchie des objets de discours en distinguant les objets de discours principaux et les objets de discours secondaires. « Un objet de discours est principal ou subordonné par rapport à un autre en fonction de la place occupée dans la structure hiérarchique par l'acte qui l'active. » (Roulet, 1999 : 61). Le couplage de la structure conceptuelle et de l'organisation informationnelle rend compte de la continuité de l'information et de la dérivation des objet de discours. « Un objet de discours qui domine un autre objet de discours dans la structure conceptuelle est premier par rapport à celui-ci et celui-ci est dérivé du premier. » (Roulet, 1999 : 62). Ces couplages assurent une analyse statique des relations entre les objets de discours, analyse qu'on peut compléter par une analyse dynamique en faisant intervenir le processus de négociation qui montre l'enchaînement des objet de discours dans le développement d'une interaction.

Le couplage entre la structure hiérarchique et l'organisation informationnelle rend compte aussi des enchaînements à distance. Il est évident que, même s'il sont séparés par 51 actes, les objets de discours *3. Les mathématiques* et *53. Les mathématiques* sont proches au plan hiérarchique.

Si on met ensemble la structure conceptuelle et l'organisation informationnelle on obtient le parcours informatif du fragment. Ainsi, on part de l'intention de définir *(3)* – pour A. L. – *(2)* les mathématiques *(3)* et on obtient le fait que les mathématiques sont une science *(4-5)* qui offre des satisfactions *(6)* esthétiques *(7-18)*, comme la musique *(19)* ou l'habitude de réfléchir seul *(20-26)*, mais aussi des satisfactions intellectuelles *(27-28)*, celles de *se cogner durement (29-30)* ou de *se battre avec son esprit (41-52).*

Le même couplage (structure conceptuelle + structure hiérarchique) rend compte des propos dérivés[18]. Il suffit de suivre le schéma conceptuel de noyaux mère au noyaux fils pour trouver l'objet de discours qui est dérivé d'un autre aux différents degrés. Par exemple, « discours cohérent » et « science hors de la science » sont des objets de discours dérivés de « mathématiques ».

L'analyse statique nous permet d'établir l'importance des objets de discours proches ou lointains l'un par rapport à l'autre dans le texte à l'aide de la position que ceux-ci occupent dans la structure hiérarchique, ainsi que la parenté des objets de discours activés par des actes qui correspondent aux concepts voisins dans la structure conceptuelle.

L'analyse dynamique explique l'émergence des objets de discours dans le processus de négociation. L'objet de discours principal *les mathématiques* est introduit par le locuteur qui a comme but la construction d'une définition des mathématiques. La réponse qu'il obtient dans un premier temps n'est pas satisfaisante, fait qui entraîne l'ouverture d'un nouvel échange, un échange de clarification qui correspond à la négociation sur *esthétiques (7)*, objet de discours réactivé par le locuteur, mais qui avait déjà été activé par l'interlocuteur *(6).* Cette négociation n'aboutit pas à l'accord des deux participants au dialogue qu'à travers deux négociations secondaires qui doivent clarifier elles aussi des objets de discours introduits par l'interlocuteur *sensibilité musicale (15)* et *réfléchir seul(21)* et réactivés par le locuteur dans les interventions qui ouvrent les négociations sur *la musique (16)* et, respectivement, sur *réfléchir seul (23).*

Le processus de négociation est développé par les deux participants au dialogue ; les objets de discours, introduits toujours par l'interlocuteur, sauf « les mathématiques », sont réactivés par

[18] Règle de couplage : « Si on a deux objets de discours OD1 et OD2, et si le concept lié à OD1 est directement dominé par le concept lié à OD2 dans la structure conceptuelle, alors OD1 est un objet de discours dérivé par rapport à OD2. » (Roulet, 2000 : 244-245)

le locuteur ou par l'interlocuteur même. Celui-ci reprend soit l'objet de discours introduit par lui- même : *intellectuel* (introduit dans l'acte *(6)* et repris dans l'acte *(28)*, soit l'objet de discours introduit par le locuteur : *les mathématiques* (activé dans l'acte *(1)* et réactivé dans l'acte *(53)*). A. L. a ainsi la possibilité de faire des commentaires et de revenir sur ses réponses pour ajouter des informations supplémentaires.

La négociation sur *honnêteté intellectuelle*, négociation initiée par A. L. se prolonge avec deux négociations secondaires sur *se cogner durement* et *se battre avec son esprit*. Ces objets de discours sont réactivés par le locuteur pour clarifier les concepts sur lesquels ceux-ci s'appuient, concepts que l'interlocuteur avait introduit dans les réponses précédentes. Chacune de ces deux négociations continue avec des négociations argumentatives qui portent sur *pourquoi* – question essentielle pour le développement des connaissances humaines.

On constate que la dynamique de l'enchaînement et de l'introduction des objets de discours est assurée par une co-participation des deux locuteurs. On pourrait distinguer trois niveaux principaux de la négociation :

– le premier niveau, sur les mathématiques dont l'objet de discours est introduit par J.N. (1) et repris par A. L. (53) ;
– le deuxième niveau où les objets de discours sont introduis par A. L. : *esthétique* et *intellectuelle* et repris par J. N. et, respectivement par A. L. ;
– le troisième niveau où les objets de discours (*musique, réfléchir seul, se cogner durement, se battre avec son esprit*) sont introduits par A. L. et repris par J. N.

6. Conclusions

L'analyse qu'on a faite assure l'interprétation efficiente du discours et assure la « contextualisation » des termes scientifiques.

En suivant le processus de négociation, on a constaté la dynamique de la co-construction du discours. La structure hiérarchique relève d'une symétrie qui permet l'anticipation de la continuité informative, fait qui va de pair avec la co-construction dynamique.

La cohésion informative est assurée par la combinaison de différents types de progression : linéaire, à topique constant et à distance. Le rôle de chacune est important dans la construction de la définition du concept – terme *mathématiques* pour A. L. La progression linéaire prolonge la chaîne informative, la progression à topique constant apporte des précisions et la progression à distance reprend les points principaux du discours.

La définition se construit autour du concept central les ***mathématiques*** par dérivation minimisante, du général au particulier, par une descendance linaire en profondeur des propriétés des mathématiques, à savoir, des concepts emboîtés : *science, satisfactions esthétiques et in intellectuelles, musique, lien avec l'enfance, se cogner, se battre avec son esprit.*

Les propos, excepté le premier, sont introduits par l'interlocuteur, le rôle du locuteur étant restreint à l'ouverture des échanges de clarification qui reprennent les objets de discours activés par le mathématicien.

Le modèle de l'analyse modulaire nous permet de préciser ce qui est commun et ce qui distingue le discours scientifique de vulgarisation d'autres discours. Ainsi, il relève d'une hiérarchisation géométrique et d'une co-construction dynamique- participative. Il est argumentatif et vise à bien clarifier tous les concepts sur lesquels sa construction s'appuie. La continuité informative n'empêche pas l'existence d'îlots informationnels relativement indépendants.

L'analyse de ce fragment met bien en évidence le caractère rigoureux et l'effort conscient de la construction d'un discours scientifique, même s'il s'agit d'un discours scientifique de vulgarisation. Elle constitue un point de départ pour un travail d'observation et de réflexion sur l'organisation de différents discours scientifiques.

Dans une perspective didactique, comprendre la structure d'un discours scientifique présente un double intérêt : d'une part, elle assure l'apprentissage des connaissances existantes, d'autre part, elle permet de mieux exprimer nos propres connaissances.

Références

Beacco J.-C. et Moirand S. (Coord.) (1995). *Les enjeux des discours spécialisés - Les carnets du Cediscor* 3. Presses de la Sorbonne Nouvelle.

Bublitz W. *et al.* (1999). *Coherence in spoken and written discourse: how to create and how to describe it*, workshop. John Benjamins.

Cicurel F., Lebre M. et Petitot G. (1993). *Discours d'enseignement et discours médiatiques – pour une rechrche de la didacticité – Les carnets du Cediscor 2*. Presse de la Sorbonne Nouvelle.

Cusin-Berche F. (Coord.) (2000). *Rencontres discursives entre sciences et politique dans les médias - Les carnets de Cediscor 6*. Presses de la Sorbonne Nouvelle.

Grobet A. (2000). *L'identification des topiques dans les dialogues*. Genève, Thèse no 472.

Grobet A. (1999). La continuité topicale dans des dialogues radiophoniques : quelques relations de discours. *Cahiers de linguistique française*, vol. (21) : 101-120.

Grobet A. (1999). L'organisation topicale de la narration. Les interactions de l'organisation topicale et des organisations séquentielle et compositionnelle. *Cahiers de linguistique française*, vol. (21) : 329-368.

Jacobi D.(1987). *Textes et images de la vulgarisation scientifique*. Peter Lang.

Jacobi D. (1999). *La communication scientifique: discours, figures, modèles*. PUG.

Kocourek R. (1991). *La langue française de la technique et de la science*. 2eedition. Oscar Brandstetten Verlag GMBH&Co.

Halliday M.A.K. et Martin J.R. (1993). *Writing Science: Literacy and Discursive Power*. Falmer (Critical Perspectives on Literacy and Education).

Martin J.R. et Veel R. (1998). *Reading Science: Critical and Functional Perspectives on Discourses of Science*. Routledge.

Nimier J. (1989). *Entretiens avec des mathématiciens. (L'heuristique mathématique).* IREM, Académie de Lyon : 15-17.

Roulet E. *et al.* (1985). *L'articulation du discours en français contemporain*. Peter Lang.

Roulet E. (1999). *La description de l'organisation du discours. Du dialogue au texte*. Didier.

Roulet E. (2000). Une approche modulaire de la complexité de l'organisation du discours. In Nølke H. et Adam J.-M. (Eds), *Approches modulaires : de la langue au discours*. Delachaux & Niestlé : 187-258.

Roulet E. *et al.* (2001). *Un modèle et un instrument d'analyse de l'organisation du discours*. Lang.

Lessico dell'intervista, lessico degli intervistati : l'articolazione tra domande e risposte nell'analisi di Alceste

Carlo Tomasetto[1], Patrizia Selleri[2]

[1] Università di Bologna (Facoltà di Psicologia) – Piazza A. Moro, 90 – 47023 Cesena (FC) – Italia
tomasetto@psice.unibo.it

[2] Università di Bologna (Facoltà di Psicologia, Dipartimento di Scienze dell'Educazione) – via Zamboni, 34 – 40126 Bologna – Italia

Abstract

Statistical analysis of responses to narrative or semi-structured open interviews challenge researchers either from a statistical or from a linguistic point of view. The dialogical aspects of interviews are often faced as threats to the analysis validity ; Instead of looking for strategies addressed to neutralize these threats, we suggest to emphasize the interactive dimension of the context in which these corpora are produced. In particular we propose some strategy to detect the effects of the interviewer-interviewed interaction by means of Alceste software.

Riassunto

L'analisi statistica di *corpora* testuali composti dalle risposte ad interviste aperte, narrative o semi-strutturate pone diversi problemi sia dal punto di vista statistico che linguistico. Anziché proporre strategie per neutralizzare le minacce alla validità dell'analisi derivate dalla natura dialogica del corpus, qui proponiamo di individuare strategie di analisi che mettano in evidenza il contesto relazionale in cui il testo è stato prodotto. In particolare proporremo alcune strategie per individuare gli effetti dell'interazione intervistatore-intervistato attraverso l'uso del software Alceste.

Parole-chiave : software Alceste, interviste semi-strutturate, analisi statistica di dati testuali, interazione, contesto discorsivo.

1. Introduzione : quando il corpus nasce da un'intervista

Le risposte a domande aperte, vuoi proposte come illustrazione di questionari strutturati, vuoi tratte da interviste o focus-group, costituiscono ormai un settore di elezione per le applicazioni della statistica testuale. Eppure, analizzare un corpus testuale ottenuto dalle risposte di molti intervistati pone una serie di problemi non superficiali (Della Ratta, 2000) ; Lebart e Salem (1988) ammettono ad esempio che questo tipo di testi può facilmente disorientare "*à la fois les statisticiens et les spécialistes des études textuelles*" (Salem, 1988 : 9) : i primi, per il carattere "*imprécis et multiforme*" del materiale a disposizione, i secondi soprattutto per l'artificiosità del contesto e per la ridondanza (altrettanto artificiale) del vocabolario (ibid.), dal momento che tutti i soggetti hanno dovuto forzatamente confrontarsi con le stesse domande.

Ad incrementare il disorientamento interviene poi l'acquisizione, ormai condivisa sul piano metodologico, delle implicazioni relazionali dell'intervista stessa. L'intervistato non è un passivo portatore di informazioni e l'intervistatore non ne è il collettore neutrale (Gubrium e Holstein, 2002), quale che sia la strategia poi impiegata in fase di analisi dei dati : elaborare i testi raccolti nel corso di un'intervista senza prendere in considerazione il suo carattere di attività conversazionale situata (Atkinson e Coffrey, 2002) rischia di minare considerevol-

mente la validità delle conclusioni che se ne traggono, e questo ha giustificato in anni recenti critiche radicali all'utilizzo dell'intervista stessa come metodo di ricerca sociale. Senza arrivare a tali estremi, è stato comunque sottolineato che il materiale testuale proveniente dalle interviste libere va trattato con estrema cautela, e che uno dei problemi ai quali porre la maggiore attenzione è proprio quello del ruolo e dell'influenza esercitata dall'intervistatore nel corso dell'interazione (Hammersley, 2003).

Un corpus raccolto attraverso una o più interviste emerge, di fatto, nel corso di interazioni nelle quali entrambi i partecipanti negoziano la loro reciproca posizione (Py e Grossen, 1997), con gli effetti pragmatici e linguistici che tutto ciò comporta. Proprio perché un'analisi statistica si basa sul lessico delle risposte date dagli intervistati, dobbiamo tenere presente che anche l'intervistatore interviene nella gestione dell'interazione discorsiva : introduce i temi, chiede spiegazioni ed esempi, riformula concetti per rilanciare un discorso interrotto e così via. Soprattutto, l'intervistatore si serve di un lessico particolare nella formulazione stessa delle domande ed è con tali scelte lessicali che l'intervistato si confronta nel momento in cui costruisce il proprio contributo al mantenimento dell'interazione.

Per quanto Lebart e Salem riconoscano nell''artificiosità del contesto' uno dei problemi sollevati dalle risposte a domande aperte, difficilmente ci si è posti il problema di far emergere da un'analisi statistica la componente dialogica di questi testi, mettendo in risalto i particolari meccanismi di interazione conversazionale che ne orientano la produzione. L'interesse non è naturalmente quello di utilizzare software e procedure di trattamento statistico per proporre un'analisi conversazionale, studiare gli scambi e le prese di turno, né studiare esaustivamente l'intervista in quanto pratica sociale. D'altro canto, non consideriamo soddisfacente nemmeno la neutralizzazione della dimensione relazionale, vista come minaccia alla validità dell'analisi statistica dei testi. Riteniamo invece che le tracce dello scambio in cui gli intervistati si sono impegnati siano un elemento con cui confrontarsi consapevolmente, sforzandosi (se possibile) di farne emergere le implicazioni per il lavoro interpretativo del ricercatore.

Rinunciando ad ogni pretesa di esaustività, affronteremo l'argomento limitandoci alle procedure e ai risultati offerti dal software ALCESTE (Reinert, 1986 e 1998), coscienti che la riflessione potrebbe essere estesa a procedure e software diversi.

L'analisi delle interviste con ALCESTE

Applicando una procedura di Classificazione Discendente Gerarchica, ALCESTE raggruppa gli enunciati di un corpus testuale in modo da creare classi il più possibile omogenee al loro interno (cioè composte da enunciati in cui si ripete sistematicamente la co-occorrenza delle stesse parole), ed il più possibile diverse l'una dall'altra. Procedendo nell'analisi, ALCESTE determina il vocabolario specifico di ciascuna classe, composto dalle parole significativamente più presenti negli enunciati della classe rispetto a tutto il resto del corpus. L'unità di analisi principale è quindi l'enunciato, al quale solo in un secondo tempo si abbina l'analisi delle singole parole. Per questo l'autore del software definisce il suo metodo più vicino alle procedure di analisi del contenuto che agli strumenti lessicometrici (Reinert, 1997). A differenza delle tradizionali analisi del contenuto, però, ALCESTE propone un'analisi della struttura interna di un corpus che non è determinata dalle griglie di post-codifica decise dal ricercatore, ma scaturisce dalla distribuzione delle parole scelte dai locutori per parlare di un certo argomento. Per riprendere la definizione di Reinert (1993), le classi omogenee di enunciati rappresentano la traccia osservabile di "*mondi lessicali*", ovvero di repertori lessicali ai quali i locutori attingono, più o meno consapevolmente, nel discutere di un certo argomento.

Da un punto di vista linguistico e statistico sono state però espresse perplessità sull'impiego di ALCESTE nell'analisi di testi compositi, nati dall'accorpamento delle risposte di soggetti

diversi (Della Ratta, 2000), non essendo a priori possibile sapere se locutori diversi possano considerarsi portatori degli stessi mondi lessicali.

L'opinione di Reinert è naturalmente diversa (Fenoglio e Reinert, 1998), sia dal punto di vista statistico, sia da quello linguistico. Le tracce dei mondi lessicali, infatti, non sono di per sé disponibili ed evidenti sulla superficie del testo, ma possono emergere soltanto in virtù della ripetizione sistematica delle stesse tracce lessicali in contesti di enunciazione diversi : tali regolarità si possono quindi cogliere solo su base statistica (Renert, 2001), e diventano tanto più informative quanto più si riesce a farle emergere dalla produzioni linguistiche di soggetti diversi. Se poi si considera l'atto linguistico come il prodotto di regole socialmente costruite, anonime e condivise da una comunità di locutori (Aubert-Lotarski e Capdevieille-Mougnibas, 2002), ad essere al centro dell'attenzione non è più soltanto l'individuo che produce un corpus testuale, ma la sua inserzione sociale in una comunità linguistica e in determinati gruppi sociali. Non a caso, lo stesso Reinert definisce i mondi lessicali come luoghi che stanno a metà tra "*les représentations individuelles et les pre-construits sociaux*" (Reinert, 1993 : 12). Interrogare più soggetti sugli stessi temi è dunque una strategia appropriata se si vuole far emergere la condivisione sociale dei mondi lessicali.

Tuttavia il principio secondo cui i mondi lessicali si impongono come fondamento topico dei discorsi, indipendentemente dalla volontà dei locutori, orientandone le scelte linguistiche verso alcune parole e non verso altre (Reinert, 2001), può essere condivisibile soltanto quando noi analizziamo discorsi spontanei (opere letterarie, produzioni scritte, discorsi pubblici…), o al massimo nel caso di interviste narrative nelle quali l'intervistatore assuma un ruolo volutamente periferico. Quando il corpus analizzato è invece composto dalle risposte a interviste semi-strutturate, nelle quali le sollecitazioni dell'intervistatore rivestono un peso innegabile, appare semplicistico attribuire soltanto agli intervistati la responsabilità delle scelte lessicali, senza prendere in considerazione il contesto conversazionale nel quale il corpus è stato raccolto.

Nella seconda parte di questo contributo presenteremo brevemente il contenuto delle classi emerse da un'analisi di interviste realizzate con insegnanti italiani, sforzandoci poi di trovare, tra gli indizi forniti dal programma, le tracce dell'interazione discorsiva tra gli insegnanti stessi e l'intervistatore.

2. Un esempio di ricerca : tracce di interazione nelle interviste ad un campione di insegnanti

Il corpus qui descritto è composto dalle risposte di 29 insegnanti italiani a interviste semi-strutturate di durata variabile tra i 45 e i 90 minuti. Le domande vertevano sulle riforme scolastiche promosse dal Ministero dell'Istruzione nel periodo 1997-2000[1], e lo scopo principale della ricerca era quello di identificare, nel discorso prodotto da un campione di insegnanti, l'articolazione tra le novità delle riforme scolastiche e il cambiamento nelle pratiche didattiche quotidiane.

La tabella 1 illustra brevemente l'andamento della Classificazione Discendente Gerarchica e il vocabolario specifico delle classi.

[1] I risultati dettagliati della ricerca sono in fase di pubblicazione sulla rivista *Psicologia dell'Educazione e della Formazione*.

Corpus intero

χ^2	**Classe 1** (261 u.c.e.)	χ^2	**Classe 2** (440 u.c.e.)	χ^2	**Classe 3** (101 u.c.e.)	χ^2	**Classe 4** (115 u.c.e.)
70.35	Scuola	40.76	Bambini	50.64	Orari	65.83	Riunioni
20.43	Nuovo	17.83	Insegnante/insegnare	48.17	Tradizionale	53.65	Organizzazione
16.30	Modello	15.43	Cambiare	47.01	Lezioni	49.04	Organizzativo
12.64	Perplessi	14.81	Difficoltà	40.16	Minuti	24.28	Finanziari
12.47	Autonomia	14.42	Ragazzi	33.71	Quadrimestre	20.86	Elaborare
10.10	Privata	12.65	Genitori	24.41	Flessibilità_oraria	20.86	Svolgere
8.93	Azienda	12.26	Alunni	18.40	Recupero	20.86	Fondi
8.70	Cambiamenti	7.15	Responsabilità	17.53	Scelte	20.51	Progetti
8.58	Politica	7.09	Colleghi	16.78	Pomeriggi	9.40	Consigli_d'Istituto
6.56	Idea	6.55	Dialogo	16.78	Sperimentale	8.44	Problemi
6.40	Pubblica	5.45	Spiegazione	15.68	Decisioni	5.14	Risolvere
6.40	Contrari	4.43	Fatica	9.51	Bimestre	5.14	Costi
4.33	Tradizione	4.02	Ascolto	6.23	Interrogazioni	4.79	Iniziative
4.27	Favorevoli	3.59	Classe	4.90	Noi	4.09	Attività

Tabella 1. Vocabolario specifico delle classi. Il corpus conta 44.465 occorrenze, con 5.702 forme distinte ; delle 1171 u.c.e. inizialmente estratte, 917 (pari al 78.31%) risultano stabilmente classificate.

L'analisi produce in primo luogo una completa scissione tra due piani discorsivi : quello legato alle riforme e al cambiamento scolastico (classi 1 e 2), e quello sulla gestione organizzativa della vita della scuola (classi 3 e 4). La prima macro-classe comprende a sua volta due nuclei lessicali ben distinti tra loro. La classe 1 riguarda le riforme propriamente dette, con enunciati che pongono in primo piano l'aspetto ideologico e politico del dibattito sulle riforme varate negli ultimi anni[2]. Contrapposto al lessico sulle riforme, nella classe 2 troviamo l'ambito discorsivo centrato sul cambiamento nelle pratiche di insegnamento. Il contesto di riferimento non è più quello socio-politico, ma quello più strettamente interpersonale : le figure di riferimento fanno tutte parte del contesto interpersonale più immediato del docente (colleghi, dirigenti, genitori) ancorando il discorso al mondo delle interazioni faccia-a-faccia che caratterizzano la vita della singola scuola.

Il secondo dei due macro-contesti fa riferimento all'organizzazione della vita quotidiana della scuola. Troviamo in questo caso enunciati relativi alla scansione temporale delle attività di routine di ogni scuola (classe 3), oppure riferimenti a compiti gestionali e burocratici, tra i quali l'impiego delle esigue risorse finanziarie (classe 4). Tutto questo, però, senza espliciti riferimenti lessicali alle riforme che hanno modificato la gestione delle scuole, né alle ripercussioni didattiche delle decisioni organizzative.

Dal profilo delle classi al lessico dell'interazione

Dopo aver descritto brevemente il vocabolario specifico e l'articolazione tra le classi possiamo chiederci : quanto ha inciso l'interazione con l'intervistatore nel far emergere proprio queste classi, e nello stabilire proprio questi rapporti lessicali tra i diversi ambiti discorsivi?

Naturalmente le domande dell'intervistatore non vengono prese in considerazione nell'analisi ALCESTE. Nel caso di interviste fortemente strutturate, o di domande aperte inserite in un

[2] Tra le riforme prese in considerazione nel corso dell'intervista ricordiamo : l'autonomia scolastica (ampliamento delle competenze gestionali e didattiche delle singole scuole), il riordino dei cicli (accorpamento della scuola elementare e secondaria inferiore in un unico ciclo di base di 7 anni), il nuovo obbligo scolastico e formativo (innalzamento della frequenza scolastica a 15 anni e della formazione obbligatoria a 18 anni).

questionario standardizzato, si può operazionalizzare l'interazione tra risposte e domande trasformando queste ultime in variabili illustrative (ad esempio codificandole come domanda 1, domanda 2, ecc.) : in questo modo è relativamente semplice andare a verificare se determinate ricorrenze lessicali siano o meno associate al contenuto specifico di una o più domande. Tra l'altro, ALCESTE offre diverse possibilità di analisi supplementari (Analisi Fattoriale delle Corrispondenze, Classificazioni Discendenti Gerarchiche *tri-croisées*, ecc.) attraverso le quali le variabili-domande potrebbero essere sistematicamente poste in relazione con il profilo lessicale delle risposte ottenute.

Nel caso di interviste più aperte, narrative o semi-strutturate, questa strada non è più percorribile, perché l'intervistatore, oltre a introdurre nella conversazione una serie di temi predefiniti, interviene con sollecitazioni ("*mi può spiegare meglio cosa intende con questo…?*", "*può farmi un esempio?*", ecc.) o riformulazioni ("*quindi Lei intende dire che…?*", ecc.), e allo stesso modo l'intervistato può condurre la sua narrazione verso temi del tutto originali rispetto a quelli previsti nella griglia dell'intervista. In questo caso, gli interventi dell' intervistatore non possono essere categorizzati in variabili illustrative, data la loro imprevedibilità, né possono entrare a far parte del corpus testuale analizzato dal software, anche perché la loro ripetitività si rivelerebbe del tutto ridondante dal punto di vista lessicale.

Ciò non toglie che l'esigenza di ricondurre l'output statistico di ALCESTE ad una dimensione più ecologica e interattiva, che tenga conto dei processi di co-costruzione del testo nel corso dell'intervista, sia altrettanto (se non maggiormente) condivisibile a mano a mano che l'interazione discorsiva si fa più libera e imprevedibile. Quelle che proponiamo qui di seguito non sono che due tra le strade percorribili quando il corpus proviene da interviste poco direttive : si tratta in entrambi i casi di piste esploratorie, che non necessitano di ulteriori interventi o manipolazioni e passano semplicemente attraverso un'attenta revisione dei materiali già a disposizione del ricercatore interessato.

La prima strada percorribile è quella di un esame indiretto, a posteriori, del materiale testuale analizzato : una volta descritte le classi prodotte dalla Classificazione Discendente Gerarchica, possiamo riesaminarne i contenuti e l'articolazione interna facendo riferimento alla griglia utilizzata per condurre l'intervista. Non crediamo possa creare scandalo il ricorso a inferenze extra-testuali per guidare il lavoro interpretativo ; è del resto risaputo che il trattamento statistico dei testi implica molti successivi passaggi (a volte frutto di tentativi ed errori) prima di approdare alla preparazione del corpus, alla sua elaborazione numerica e all'interpretazione dei risultati ottenuti (Bolasco, 1999).

Rileggendo le classi alla luce della griglia utilizzata per guidare l'intervista sembra emergere chiaramente che tutte le richieste di pareri e commenti esplicitamente legati alle riforme e ai cambiamenti didattici hanno dato luogo a risposte poi confluite nelle classi 1 e 2, mentre le numerose richieste di esemplificazione legate ad aspetti logistici e organizzativi della vita scolastica hanno dato luogo ad enunciati isolati poi nelle classi 3 e 4. Prima di imputare la scissione delle classi alle scelte lessicali degli intervistati, dobbiamo chiederci in che modo tali scelte siano maturate ; in effetti eravamo noi a sollecitare esemplificazioni riferite alle routines organizzative, perché era nostra convinzione che l'impianto legislativo delle riforme avrebbe comportato ripercussioni concrete ed osservabili proprio a questo livello del funzionamento scolastico.

Il percorso di delle classi mette in luce il fatto che gli insegnanti, almeno a livello lessicale, non hanno costruito lo stesso legame tra i due piani concettuali. Di fronte alle nostre sollecitazioni, gli intervistati hanno enumerato incombenze e scadenze della gestione quotidiana della scuola, ma senza attivare legami lessicali con i temi della riforma e del cambiamento. Se lo avessero fatto avremmo potuto reperire enunciati in cui gli intervistati citavano sia parole relative alle riforme, sia parole relative alle routines gestionali. Semplificando, avremmo

potuto incontrare enunciati come "*con l'autonomia abbiamo modificato gli orari scolastici*", oppure "*con l'autonomia non abbiamo modificato gli orari scolastici*" : in entrambi i casi ALCESTE avrebbe raggruppato gli enunciati in una classe comune, vista co-occorrenza tra i riferimenti lessicali all'una e all'altra sfera discorsiva. Nel corso delle interviste, però, ciò non è avvenuto : gli insegnanti hanno ampliamente parlato tanto delle riforme quanto delle routines organizzative, ma senza stabilire alcuna contiguità tra le une e le altre all'interno degli stessi enunciati.

Lo scambio verbale che qui riportiamo mette bene in luce l'andamento conversazionale che ha portato a segmentare il discorso sul cambiamento rispetto al discorso organizzativo :

D_1 : PARLIAMO DELL'AUTONOMIA...
R_1 : Lo dicevo anche prima, ha liberato delle energie nella scuola. ad esempio ne parlavo in commissione... abbiamo concretizzato, ha liberato un protagonismo da parte degli insegnanti...
D_2 : SI ASPETTA CHE POSSA ANCORA EVOLVERE NEI PROSSIMI ANNI?
R_2 : spero di sì, perché è una grande carta secondo me, il che vuol dire che la scuola, con il POF, propone certe cose alla società... e quindi è un peccato se vogliamo. Può legarsi bene ai bisogni della società. […].
D_3 : RISPETTO ALL'ORGANIZZAZIONE DELLE LEZIONI O DEGLI ORARI AVETE MODIFICATO QUALCOSA CON L'AUTONOMIA?
R_3 : C'è qualche margine in cui si può decidere. Abbiamo fatto la flessibilità del sabato, per dire, che consiste nel chiudere la scuola alle dieci e mezza per il biennio. Adesso probabilmente ci presteremo per fare la settimana corta, senza il sabato (*matematica, superiori, oltre 50 anni, maschio, nessun incarico*).

Quadro 1. Esempio di scambio sulle conseguenze organizzative di una riforma.

Lo scambio appare in qualche misura forzato : di fronte alla suggestione sulla riforma dell' autonomia, l'insegnante esprime la sua opinione sulle opportunità di cambiamento, sulla relazione scuola-società e sulle aspettative future (D_1-R_1 e D_2-R_2) ; quando noi domandiamo se ci sono state ripercussioni organizzative e gestionali (D_3), l'insegnante risponde con eguale disponibilità, ma il discorso scivola verso un altro ambito di contenuto, che è quello dei tempi della scuola e delle relative negoziazioni (R_3) : in effetti, quella che per noi era una semplice richiesta di esemplificazione, per l'intervistato ha comportato una ridefinizione del tema della conversazione.

La distanza tra i due ambiti discorsivi si traduce, sul piano metrico e lessicale, in una netta separazione tra classi i cui enunciati non presentano riferimenti in comune. La genesi di questa separazione va riconosciuta nell'interazione tra una rappresentazione del fenomeno introdotta dall'intervistatore (le cui domande non sono mai neutre), una rappresentazione costruita dall'intervistato (che forse non aveva mai esplicitamente pensato ad un legame tra riforme e routines organizzative), e un contratto di comunicazione asimmetrico che lega i due protagonisti (l'intervistatore chiede, il soggetto risponde).

Lo stralcio riportato presuppone il ricorso a risorse diverse da quelle fornite dal software, in particolare una rilettura attenta delle trascrizioni originali delle interviste. La seconda strada che proponiamo di seguire prende invece in considerazione alcuni elementi forniti direttamente dall'output statistico di ALCESTE.

Abbiamo detto che il testo delle domande viene volutamente escluso dalle analisi, vista la sua ridondanza lessicale, e per ottenere tale esclusione è sufficiente strascrivere gli interventi dell'intervistatore in caratteri maiuscoli. Nel corso della suddivisione automatica degli enunciati può accadere che alcuni segmenti di domande vengano inclusi nelle unità di testo sottoposte ad analisi. In questo caso il testo delle domande, scritto in stampatello, non concorre all'analisi e alla formazione delle classi, ma viene ripreso in un secondo momento come

variabile supplementare : in particolare, il software calcola a posteriori la distribuzione delle parole scritte in stampatello e, nel caso tali parole compaiano più frequentemente negli enunciati di una classe rispetto al resto del corpus, provvede ad inserirle in coda al vocabolario specifico della classe. Nelle tabelle 2 e 3 sono elencate le parole in maiuscolo (cioè provenienti da spezzoni di domande) confluite rispettivamente nel vocabolario delle classi 1 e 2 ; nelle tabelle riportiamo anche l'eventuale presenza degli stessi vocaboli nel lessico delle risposte.

CLASSE 1			
Domande	**Risposte**	**Domande**	**Risposte**
ASPETTA	aspett+	PAROLA	---
AUTONOMIA	autonomi+	PROSSIMI	---
OBIETTIVI	obiettiv<	RIORDINO	---
PROPOSTO	propo<	SITUAZIONE	---
DALLA	---	STUDIATO	---
ENTRATA	---	VIGORE	---

Tabella 2. Parole utilizzate nelle domande e associate al vocabolario specifico delle classi 1 e 2 ; in corsivo le stesse parole reperite nel vocabolario specifico delle classi. Le parole provenienti dalle domande non vengono lemmatizzate da ALCESTE, dal momento che la loro attribuzione alle classi ha finalità puramente illustrative.

CLASSE 2			
Domande	**Risposte**	**Domande**	**Risposte**
ALUNNI	alunn+	RAGAZZI	ragazz<
BAMBINI	bambin<	RAGAZZINI	
CAMBIARE	cambi+	RIFERIMENTO	riferimento
CAMBIATI		BENE	---
CAMBIATO		DIRIGENTE	---
CAMBIEREBBE		FARE	---
CHIEDE	chied+	INCERTEZZA	---
CHIESTO		INTERNO	---
COLLEGA	collegh+	IPOTETICO	---
GENITORE	genitor<	LAVORARE	---
GENITORI		METODO	---
INSEGNAMENTO	insegn+ [professo<]	PICCOLI	---
INSEGNANTE		SCUOLA	---
INSEGNANTI		SCUOLE	---
INSEGNARE		SETTENNIO	---

Tabella 3. Parole utilizzate nelle domande e associate al vocabolario specifico delle classi 1 e 2 ; in corsivo le stesse parole reperite nel vocabolario specifico delle classi.

Limitiamo l'esemplificazione alle prime due classi. È evidente che parte del lessico utilizzato dagli intervistati è stato introdotto dal ricercatore attraverso la formulazione delle domande, anche se la ripetizione di forme identiche si limita tutto sommato a pochi vocaboli.

Per quanto utile, l'esame dei frammenti di domande non fornisce indizi del tutto esaustivi, dal momento che l'inclusione di tali frammenti negli enunciati analizzati riveste carattere del tutto aleatorio[3]. Non solo, ma fermarsi al confronto del lessico delle domande con quello delle

[3] ALCESTE offre la possibilità di introdurre le parole in maiuscolo come variabili supplementari nell'Analisi Fattoriale delle Corrispondenze, rendendo così visibile e quantificabile il legame tra lessico delle domande (o almeno di quella porzione di domande incidentalmente inclusa nell'analisi) e lessico del corpus analizzato. Si tratta di una possibilità di approfondimento potenzialmente molto promettente, che però al momento incontra almeno due limiti nella sua applicazione. In primo luogo, come detto, l'inclusione di spezzoni di domande nelle u.c.e. riveste carattere alquanto aleatorio, e non si presta quindi ad un'analisi sistematica e del tutto affidabile. In secondo luogo, la visualizzazione dei piani fattoriali attualmente permessa da software non appare sufficiente-

risposte può indurre a stabilire nessi causali non corretti, e comunque incompatibili con una visione dell'intervista come co-costruzione situata : così come non vogliamo ignorare il ruolo attivo dell'intervistatore, allo stesso modo non possiamo immaginarci un intervistato passivo e manipolato nelle sue risposte.

Se, anziché limitare il confronto alle singole parole, spostiamo il confronto al livello degli enunciati, che costituiscono a tutti gli effetti l'unità di analisi di ALCESTE, scopriamo che la stessa domanda dell'intervistatore può attivare scelte lessicali del tutto diverse :

U.C.E. 317 : SECONDO LEI, NEL PROSSIMO FUTURO, IL RUOLO DELl' INSEGNANTE È DESTINATO A CAMBIARE? sì, l' #insegnante deve diventare un tuttologo e deve diventare #sempre #di-più un #riferimento per la vita quotidiana dell' #alunno, visto che ormai tutti i #genitori lavorano. (*matematica, medie, 36-50 anni, femmina, collaboratrice del dirigente ; u.c.e. della CLASSE 2*).

U.C.E. 1018 : SECONDO LEI IL RUOLO DELL' INSEGNANTE È DESTINATO A CAMBIARE? #secondo-me bisogna #vedere #qual è il progetto della #scuola #che #viene fuori, perché se un tempo il ruolo #sociale dell'insegnante era un ruolo #comunque #riconosciuto… (*lettere, superiori, 36-50 anni, femmina, nessun incarico ; u.c.e. della CLASSE 1*).

Quadro 2. Esempi di scambio ; u.c.e. estratte a partire dalla stessa domanda.

Entrambe le risposte riprendono un vocabolo presente nella domanda (*insegnante*), ma lo sviluppo che ne segue va in direzioni del tutto diverse, attivando (nei termini di ALCESTE) due mondi lessicali distinti. Nel primo caso l'insieme dei riferimenti lessicali richiama il mondo delle relazioni interpersonali che ruotano intorno alla scuola e nelle quali è immerso l'"insegnante" evocato dall'intervistatore ("*genitori*", "*alunno*", ecc.) ; in effetti, la risposta confluisce negli enunciati che compongono la classe 2. Nel secondo caso la prospettiva evocata non è più quella delle relazioni immediate, ma è quella che potremmo definire sovra-ordinata e che dall'"insegnante" arriva a comprendere la "*scuola*" e la "*società*" nel suo complesso. La risposta, in questo caso, va naturalmente a confluire tra gli enunciati della classe 1.

3. Conclusioni

Utilizzando un software statistico (nel nostro caso ALCESTE) per analizzare testi per loro natura dialogici, quali sono quelli prodotti nel corso di una o più interviste, ci siamo chiesti quali cautele dovessero essere prese per rendere conto, almeno in parte, della dimensione ecologica del corpus raccolto, soprattutto per quel che concerne le sue condizioni di produzione. Per sua natura, l'intervista costituisce una pratica discorsiva che coinvolge almeno due soggetti, ciascuno dei quali è attivo costruttore di conoscenza nell'ambito di uno specifico contratto di comunicazione ; l'attività di co-costruzione nella quale i due attori sono impegnati si manifesta a livello pragmatico, contenutistico ed anche sul piano prettamente lessicale, incidendo sui risultati che l'elaborazione statistica del testo fornisce alla nostra interpretazione.

La prima strada che abbiamo proposto per un riesame critico dei risultati di ALCESTE prevede il ritorno al testo originario, confrontando criticamente la segmentazione degli enunciati ottenuta attraverso la Classificazione Discendente Gerarchica, con la scansione dei temi proposti dal ricercatore nel corso dell'intervista. Il secondo percorso di revisione prende invece in considerazione indizi presenti nell'output statistico fornito dal software, ed in particolare si concentra sulle eventuali corrispondenze tra lessico impiegato nella formulazione delle

mente chiara e richiede un cospicuo lavoro di ripulitura da parte dell'utente. L'AFC ci sembra comunque una delle strade più promettenti in vista di un più rigoroso esame su base statistica delle interazioni discorsive.

domande (inserito a posteriori nei risultati come elemento illustrativo supplementare), e lessico utilizzato dagli intervistati nelle loro risposte. Questa seconda opzione presenta elementi estremamente aleatori e deve essere senz'altro raffinata ; ciò non di meno entrambe le strade consentono di individuare alcune forzature nelle scelte lessicali degli intervistati.

Il legame tra domande e risposte non va visto nè in termini neutrali, nè in termini deterministici : non ci sembra corretto analizzare un corpus di interviste ignorando o sopprimendo *tout court* il ruolo dell'intervistatore, né immaginare che le risposte degli intervistati seguano pedissequamente le suggestioni delle domande. Assumere una prospettiva dialogica e interazionista permette non solo di studiarne l'organizzazione interna di un testo analizzato su base statistica, ma soprattutto di individuare i punti in cui questa organizzazione si discosta da quella immaginata a priori (e implicitamente veicolata) dal ricercatore. Un interessante sviluppo della ricerca in questa direzione potrebbe partire dal confronto sistematico (nella forma, nella struttura e nel contenuto) di *corpora* prodotti nel corso di interazioni dialogiche gestite in modi diversi, ad esempio comparando tra loro interviste condotte con un grado si strutturazione più o meno forte. A questo scopo, diversi approcci di indagine, a partire dai classici strumenti della lessicometria (orientati allo studio della ricchezza e della diversificazione del lessico prodotto in situazioni diverse), per arrivare agli sviluppi più recenti delle analisi discriminanti e delle funzioni di regressione, potrebbero fornire indizi importanti per un ulteriore approfondimento teorico e metodologico .

Naturalmente, quelle che abbiamo illustrato non sono che due tra le strade percorribili da chi voglia ancorare l'analisi statistica di un testo alle concrete condizioni nelle quali il testo stesso è stato prodotto. Abbiamo proposto, a titolo di esempio, due strategie che qualsiasi utente di ALCESTE può impiegare a partire da un riesame critico dell'output statistico e/o del corpus di partenza, ben sapendo che il problema affrontato meriterebbe un approfondimento ben maggiore. Un ulteriore sviluppo nelle tecniche multidimensionali per il trattamento di dati testuali potrà senz'altro offrire al ricercatore interessato strumenti di analisi più sistematici, rendendo più agevole e proficuo l'esame congiunto di un corpus testuale e dell'interazione sociale che ad esso ha dato origine.

Riferimenti bibliografici

Atkinson P. e Coffrey A. (2002). Revisiting the relationship between participant observation and interviewing. In Gubrium J.F. e Holstein J.A. (Eds), *Handbook of interview research*. Thousand Oaks, Sage : 800-814.

Aubert-Lotarski A. e Capdevieille-Mougnibas V. (2002). Dialogue méthodologique autour de l'utilisation du logiciel Alceste en sciences humaines et sociales : 'lisibilité' du corpus et interprétation des résultats. In *Actes des JADT 2002* : 45-56.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Della Ratta F. (2000). L'analisi testuale : uno strumento per la ricerca sociale. *Sociologia e ricerca sociale*, vol. (61) : 102-127.

Fenoglio E. e Reinert M. (1998). Application d'*Alceste* à un corpus d'entretiens : de l'analyse de données à l'analyse de l'énonciation. In *Actes des JADT 1998* : 223-229.

Gubrium J.F. e Holstein J.A. (2002). Introduction. In Gubrium J.F. e Holstein J.A. (Eds), *Handbook of interview research*. Thousand Oaks, Sage : 3-32.

Hammersley M. (2003). Recent radical criticism of interview studies : any implication for the sociology of education? *British Journal of Sociology of Education*, vol. (24/1) : 119-126.

Lebart L. e Salem A. (1988). *Analyse statistique des données textuelles : questions ouvertes et lexicométrie*. Dunod.

Lebart L. e Salem A. (1994). *Statistique Textuelle*. Dunod.

Py B. e Grossen M. (1997). Introduction Générale : Interactions, médiations et pratiques sociales. In Grossen M. e Py B. (sous la direction de), *Pratiques sociales et médiations symboliques*. Peter Lang.

Reinert M. (1986). Un logiciel d'analyse lexicale. *Les Cahiers de l'Analyse des Données*, vol. (4) : 471-484.

Reinert M. (1993). Les « mondes lexicaux » et leur « logique » à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et Société*, vol. (66) : 5-37.

Reinert M. (1997a). Les "mondes lexicaux" de six numéros de la revue « Le Surréalisme au service de la Révolution ». *Mélusine*, L'Age d'Homme, vol. (XVI) : 270-302.

Reinert M. (1998). Mondes lexicaux et topoi dans l'approche Alceste. In Mellet S. e Vuillaume M. (sous la direction de), *Mots chiffrés et déchiffrés*. Honoré Champion.

Reinert M. (2001). *Alceste*, une méthode statistique et sémiotique d'analyse de discours ; application aux « *Rêveries du promeneur solitaire* ». Papier présenté au *XIIème colloque de Psychiatrie de Marseille*, 17-19 Mai 2001.

Finance organisationnelle : un essai de représentation*

Stéphane Trébucq

CRECCI (Centre de Recherche en Contrôle et Comptabilité Internationale) – Université Montesquieu Bordeaux IV – 33600 Pessac – France
trebucq@u-bordeaux4.fr

Abstract

Corporate governance is one of the most stimulating research areas in the field of organizational finance. Due to the huge number of publications in this field, our work is based on several articles written by G. Charreaux. Our objective is to build some mappings showing the evolution of governance, and its main concepts. After a detailed presentation of our methodology, synthetic mappings are proposed, using several softwares : Tropes, Lexter And Decision Explorer. Results are then discussed in the third part of the article, and some ways of improvement are proposed.

Keywords : Tropes, Lexter, Decision Explorer, corporate governance, theory, representation, content analysis, semantic analysis, syntaxic analysis, ontology, mapping, concepts, actors

Résumé

Le gouvernement des entreprises constitue une des voies de recherche les plus stimulantes de la finance organisationnelle. Face à l'importance des écrits, et sur la base d'une série d'articles écrits par G. Charreaux, l'objectif de cet article est de construire une cartographie de l'état et de l'évolution des principaux concepts mobilisés. Après une présentation détaillée de la méthodologie, différentes cartes de synthèse sont proposées grâce aux logiciels Tropes, Lexter et Decision Explorer. Les résultats sont ensuite discutés dans la troisième partie et des pistes d'enrichissement finalement suggérées.

Mots-clés : Tropes, Lexter, Decision Explorer, gouvernance, gouvernement d'entreprise, théorie, représentation, analyse de contenu, analyse sémantique, analyse syntaxique, ontologie, cartographie, concepts, acteurs.

1. Introduction

Le présent article s'inscrit dans une approche cognitive, à notre connaissance inédite dans le domaine de la finance organisationnelle[1]. Pour cette raison, une première partie est consacrée à la méthodologie mise en œuvre. Les résultats de l'étude empirique présentés dans une seconde partie permettent d'aboutir à des observations longitudinales et à la constitution d'une série de cartes de synthèse. La troisième partie est consacrée à une discussion de ces cartes, ainsi qu'aux perspectives d'enrichissement restant à développer dans le cadre de futures recherches en matière de gouvernement d'entreprise.

* L'auteur remercie, pour leur aide et leurs conseils, Martine Hlady-Rispal, Pierre Molette, Laurence Bardin, Didier Bourigault, Bernard Dousset, Pierre Fiala et Serge Heiden.

[1] En ce qui concerne la France, G. Charreaux est l'un des auteurs académiques pouvant être considéré parmi les plus emblématiques d'un tel courant de recherche. Réalisé grâce aux avancées technologiques opérées dans le champ de l'analyse de contenu et de la cartographie cognitive, et en s'appuyant sur un corpus de 144 573 mots, formé à partir de 14 articles publiés par cet auteur entre 1987 et 2000, la présente communication a pour objet de parvenir à une meilleure compréhension des modes de conceptualisation et de pensée relatifs aux problèmes posés en finance organisationnelle.

2. Méthodologie d'analyse de contenu et cartographie

Comme le précisent Cossette et Audet (1994 : 15), une carte cognitive est une illustration graphique de la représentation mentale que le chercheur se fait « d'un ensemble de représentations discursives énoncées par un sujet à partir de ses propres représentations cognitives, à propos d'un objet particulier ». Sur un plan épistémologique, les chercheurs spécialistes de la cartographie cognitive revendiquent généralement un positionnement constructiviste (Cossette, 1999 : 4), la compréhension de la pensée d'autrui nécessitant non seulement l'intervention du chercheur, mais aussi la mise en œuvre d'une certaine subjectivité. Cependant, si les concepts peuvent faire l'objet d'une sélection manuelle et raisonnée, des moyens technologiques innovants autorisent désormais des approches complémentaires permettant d'identifier des concepts clés.

Au nombre de ceux-ci figurent notamment l'analyse automatique de contenu et l'analyse cognitivo-discursive (ACD)[2] développées par Ghiglione et Blanchet (1991) et Ghiglione *et al.* (1998). Par le truchement de ces techniques, on peut obtenir la liste des référents-noyaux (désormais RN) et des propositions remarquables utilisant les termes clés d'un texte. L'analyse syntaxique (LEXTER-Logiciel d'EXtraction de TERminologie) mise au point par Bourigault (2000), facilite également l'acquisition de connaissances en mettant à jour « l'organisation d'une ontologie en champs conceptuels ».

Ces différentes analyses reposent sur l'usage de logiciels spécifiques améliorant considérablement l'objectivité, la stabilité et la reproductibilité des résultats. Il est important de bien saisir l'essence de ces méthodologies, afin de mieux comprendre et interpréter les résultats obtenus. En effet, comme le rappellent Allard-Poesi *et al.* (1999), les résultats dépendent fortement des unités d'analyse retenues (mots, groupes de mots, phrases, paragraphes, texte entier), de la catégorisation et d'un éventuel regroupement de ces unités (lexicales, sémantiques, thématiques), des combinaisons de techniques quantitatives et qualitatives choisies afin d'identifier des liens, et enfin, des objectifs poursuivis par le chercheur (décrire, comparer, expliquer ou prédire).

2.1. *L'analyse sémantique*

L'objectif de la présente recherche est de parvenir à élaborer, sur la base d'une série d'articles rédigés par G. Charreaux, une cartographie rassemblant les concepts fondamentaux utilisés en matière de gouvernance d'entreprise. Le résultat obtenu n'est donc pas une cartographie exacte et précise de la pensée de l'auteur, mais plutôt un diagramme ou graphe de synthèse des concepts utilisés dans ses textes, et de leurs associations. Dans le cas de l'analyse automatique de contenu effectuée à l'aide du logiciel Tropes (version 5), l'unité d'analyse est la proposition grammaticale (sujet/verbe/complément). De l'avis de Ghiglione *et al.* (1998), « la proposition constitue une unité de découpage [du texte] qui tout en respectant la surface textuelle, permet la mise en scène langagière d'un univers minimal ». Dans le cadre de cette unité d'analyse, il est possible de repérer non seulement les termes pivots ou RN (objets ou personnes) de chaque proposition mais aussi les relations entre ceux-ci.

À titre d'exemple, dans une proposition du type « le dirigeant peut avoir un intérêt à accroître ses prélèvements non pécuniaires », ou encore « les dirigeants agissent dans les intérêts des anciens actionnaires », l'analyse automatique du logiciel Tropes sera en mesure d'identifier

[2] L'analyse cognitivo-discursive s'appuie sur une série de comptages et d'algorithmes, afin de détecter dans un texte des propositions considérées comme remarquables. Autrement dit, celles-ci peuvent être considérées comme les plus représentatives du texte étudié.

les termes clés de dirigeant et intérêt. Un classement sémantique sera ensuite constitué sous les catégories de « patron » (avec pour classe d'équivalents[3] : dirigeant, employeur, administrateur, gérants) et d'« intérêt » (avec pour classe de termes : intérêt ou taux d'intérêt). La relation[4] entre ces deux RN, observée dans les deux propositions précédentes, est symbolisée de la manière suivante : patron→intérêt. On parlera alors de cooccurrence des deux RN. La flèche qui les relie précise le sens de la relation, patron étant dans les propositions en situation d'actant, et intérêt en situation d'acté. L'action, représentée dans les propositions par les verbes « avoir » ou « agir » est alors symbolisée par une flèche.

Le logiciel Tropes propose une navigation hypertexte en offrant la possibilité de consulter et de relire non seulement les propositions mais aussi les passages du texte dans lesquels ces mises en relation sont réalisées. À l'aide de ce logiciel on peut ainsi entreprendre véritablement une lecture innovante et différente des textes, selon des modalités ne pouvant être mises en œuvre dans le cadre d'une lecture cursive classique. À cela s'ajoute la possibilité d'explorer la dimension longitudinale et historique, le corpus de textes étudiés ayant été rédigé sur la période 1987-2000. Les outils de comptage et leur représentation en histogramme permettent d'observer au fil du temps la permanence ou les fluctuations des RN utilisés, les éventuelles introductions de termes nouveaux, ainsi que des mises en relation inédites[5]. À titre d'exemple, de nouvelles expressions sont apparues récemment telles que « création de valeur » à partir de 1996, et « valeur partenariale » ou « partage de la valeur » à partir de 1998.

Ces résultats présentent l'avantage d'ouvrir « un espace d'interrogation sur le texte » (Ghiglione *et al.*, 1998 : 47) fondé sur la fréquence d'apparition inattendue de certains RN, ainsi que sur leur place en situation d'actant ou d'acté. En cas de nécessité, on peut envisager de construire des classes de RN spécifiques en élaborant un « scénario »[6] approprié. En définitive, cela revient à élaborer un filtre appliqué à l'ensemble du corpus textuel sélectionné, exclusivement en fonction des classes de RN définies dans le « scénario ».

La mise en pratique de l'analyse sémantique, réalisée à l'aide du logiciel Tropes, n'est pas sans présenter quelques limites dans la mesure où l'on recourt aux seuls traitements automatisés. C'est ainsi que la liste des mises en relations opérées au sein de l'univers de référence 1 aboutit à des résultats peu lisibles, du fait des classes d'équivalents souvent trop extensives lorsqu'il s'agit d'analyser de façon pertinente le vocabulaire de la gouvernance. À contrario, au niveau d'agrégation le plus faible, la mise en relation des références utilisées aboutit fréquemment à identifier des syntagmes[7] nominaux. Ainsi, les mises en relation « théorie→agence », « prix→agence », »prix→transaction », « relation→agence » ou encore

[3] Les classes d'équivalents regroupent les mots qui apparaissent fréquemment dans le texte et qui possèdent une signification très voisine.

[4] Les relations indiquent quelles sont les classes d'équivalents les plus fréquemment reliées, c'est-à-dire rencontrées côte à côte à l'intérieur d'une même proposition.

[5] Trois classes de regroupements sémantiques sont proposées, par niveau croissant d'agrégation. Ainsi, sans autre retraitement, le logiciel Tropes fournit un univers de référence 1, formé de concepts d'une grande généralité, un niveau intermédiaire, l'univers de référence 2, et enfin, les références utilisées correspondant aux classes d'équivalents les moins étendues. Les mises en relation des référents-noyaux peuvent, de ce fait, être réalisées à différents niveaux de regroupement. A titre d'exemple, à partir du corpus des articles de Charreaux, le référent-noyau principal de l'univers de référence 1 est la « finance », dans l'univers de référence 2 il s'agit de la « hiérarchie », et pour les références utilisées il s'agit du « patron » ou « dirigeant » d'entreprise.

[6] Un scénario est composé d'un certain nombre de groupes sémantiques définis par l'utilisateur du logiciel Tropes, faisant ensuite l'objet d'une analyse spécifique.

[7] Un syntagme représente un « groupe de mots qui se suivent avec un sens ». Il constitue une unité dans une organisation hiérarchisée de la phrase (définition du dictionnaire *Robert*).

« création→valeur » correspondent respectivement aux syntagmes « théorie de l'agence »[8], « coûts d'agence », « coûts de transaction », « relation d'agence » ou « création de valeur ». En l'occurrence, la flèche censée symboliser la notion d'influence entre un actant et un acté peut être mal interprétée ou pour le moins manquer de pertinence. Il s'avère donc plus utile de travailler sur les syntagmes. On peut regretter que le logiciel Tropes ne dispose pas d'un algorithme permettant d'identifier les syntagmes nominaux.

2.2. *L'analyse syntaxique*

Cependant, il est possible de pallier une telle carence du logiciel LEXTER, développé par Didier Bourigault. Il permet d'obtenir un classement des syntagmes, rangés en fonction de leur fréquence d'utilisation dans le corpus étudié. Ainsi, dans le cas des textes de Charreaux, le syntagme « conseil d'administration » est utilisé avec la plus forte fréquence, soit 193 fois. Le second : « coût d'agence » est cité 115 fois. Pour l'ensemble du corpus, le nombre total de syntagmes obtenus est considérable. Il atteint un effectif total de 978. Opérer une représentation graphique lisible de l'ensemble des mises en relation ne peut donc être raisonnablement envisagé. En revanche, un travail utile de classement des syntagmes par grandes catégories peut être réalisé. C'est ainsi, par exemple, qu'au terme de « valeur » se rattacheront des syntagmes-*actants* du type « valeur partenariale », « valeur créée », « valeur de marché », « valeur actuelle », « valeur globale », « valeurs comptables », « valeur actionnariale » et des syntagmes-*actés* du type « chaîne de valeur », « création de valeur », « destruction de valeur », « maximisation de la valeur », « partage de la valeur », « perte de valeur », « répartition de la valeur ». Par la suite, on peut sélectionner en tant que concepts-clés des têtes de syntagmes, comme par exemple le concept représenté par le mot « valeur », puis observer les relations entre ces concepts à l'aide d'un « scénario » élaboré dans Tropes.

Une représentation synthétique du réseau des relations existant entre concepts (ou têtes de syntagmes) peut être obtenue à partir du logiciel de cartographie cognitive Decision Explorer. Le formalisme graphique retenu répond à l'objectif premier de la plus grande lisibilité des cartes, le positionnement des concepts ne présentant pas d'importance particulière.

En définitive, même si les images d'un même corpus de textes sont multiples en fonction des choix méthodologiques opérés, l'utilisation complémentaire des approches sémantique et syntaxique permet, quant à elle, d'aboutir à une représentation plus fidèle, et finalement d'élaborer une cartographie inédite de textes consacrés au sujet de la gouvernance.

3. Analyses et représentations d'un corpus en gouvernance d'entreprise

Conformément à la méthodologie annoncée, les résultats des analyses sont successivement exposés selon une approche sémantique (2.1.), puis syntaxique (2.2.). L'ensemble de ces résultats a permis de procéder à une série de représentations synoptiques du corpus textuel de gouvernance étudié (2.3.).

3.1. *Analyse sémantique du corpus textuel de gouvernance*

L'analyse sémantique peut être utilisée afin d'obtenir une première représentation du contexte, et de la structuration des textes étudiés. Elle permet de mieux en appréhender l'ossature fondamentale, sans laquelle « l'édifice textuel » s'écroulerait.

[8] La relation « théorie → agence » peut aussi recouvrir des syntagmes tels que « théorie normative de l'agence » ou « théorie positive de l'agence ».

Parmi les cinquante premières références utilisées, classées par ordre décroissant de fréquence, on trouve tout d'abord le RN de « Dirigeant »[9]. Il fait l'objet du plus grand nombre de citations (1367 fois), même si on observe une chute de la fréquence d'utilisation à partir de 1995 (approche longitudinale, avec une visualisation des fréquences de citation en histogramme). D'autres RN apparaissent avec une fréquence d'utilisation élevée. Ils servent à rappeler régulièrement au lecteur le contexte de l'analyse, à savoir celui de « l'entreprise », de « l'organisation » et des « marchés », ainsi que le principal objet des articles, consacrés à l'étude des « systèmes » de gouvernement[10] et des « relations » entre les « actionnaires » et « dirigeants ». Les raisonnements s'appuient sur un ensemble de « théories », parmi lesquelles on peut noter la place privilégiée accordée à la « théorie de l'agence », et au récent regain d'intérêt qu'elle suscite (analyse des graphes longitudinaux) comme l'atteste l'article de Charreaux (1999), intitulé "La théorie positive de l'agence : lecture et relectures".

Ces premiers résultats mettent en exergue le positionnement central revenant au « dirigeant d'entreprise », même si celui-ci doit être nuancé par une fréquence d'emploi décroissante. Partant, les RN « stakeholders » (87 citations) ou « partenaires » (83 citations), incontournables dans le cadre d'une théorie élargie à l'ensemble des parties prenantes, voient croître leur fréquence d'utilisation à partir de 1995, même s'ils semblent d'une moindre importance au sein du corpus textuel pris dans sa globalité.

Cette première caractérisation peut être approfondie si l'on concentre l'effort d'analyse sur la liste des RN suivants s'échelonnant jusqu'au rang 73. Ils représentent en effet la fondation même des textes étudiés, et définissent les problématiques de gouvernance par rapport aux champs du « contrôle », de la « valeur », et de la « performance ». La pensée et les raisonnements développés en matière de gouvernance ne peuvent s'organiser sans recourir à ces concepts. À ceux-ci, on peut ajouter l'usage complémentaire de RN traitant de « mécanismes », de « transactions », de « décisions », « d'investissements », et de « risques ».

Un résumé graphique des mises en relation retenant 54 concepts (RN) et de 73 liens (flèches)[11] peut être construit à l'aide du logiciel Decision Explorer (voir figure 1).

Figure 1

[9] Le logiciel Tropes regroupe sous le référent-noyau « patron » des termes assez différents comme ceux de dirigeant et d'administrateur. Ces termes peuvent toutefois être distingués dans le cadre d'un scénario, le nombre d'utilisations s'élevant à 1229 pour le terme « dirigeant » et 155 pour le terme « administrateur ».

[10] Au sein du corpus textuel retenu, l'usage proprement dit des termes « gouvernement » ou « gouvernance » (d'entreprise) est assez récent, puisque ceux-ci n'ont été respectivement utilisés qu'à partir de 1994 et de 1998.

[11] Se reporter à la note 3, pour la définition d'une relation symbolisée par une flèche.

La cartographie ainsi obtenue résume les mises en relations les plus courantes, et met à jour les associations fondamentales observées entre les concepts de gouvernance. Après calcul des scores de centralité des concepts du graphe, l'importance accordée au dirigeant d'entreprise, aux systèmes de contrôle de celui-ci ainsi qu'aux notions de valeur et de performance est bien confirmée.

Une telle cartographie permettra à un public, désireux de mieux comprendre les sujets de gouvernance, de procéder à une prise de conscience de la variété des concepts utilisés, et de leur centralité respective. Elle constitue également une première grille d'analyse permettant de présenter la théorie de la gouvernance, et pouvant donner lieu à de nouvelles cartographies centrées sur des concepts particuliers. Elle peut, par ailleurs, servir de canevas lors d'une première lecture, ou d'une relecture, des articles de G. Charreaux, et plus généralement, constituer une base de comparaison lors de l'analyse d'autres textes passés, présents, ou futurs émanant du même auteur ou d'autres auteurs, et se rattachant au champ de la gouvernance.

3.2. *Analyse syntaxique : vers une structuration de l'ontologie de la gouvernance*

L'ingénierie des connaissances engagée à partir de l'analyse automatique de contenu peut être considérablement approfondie à partir d'une « analyse syntaxique de surface dédiée au repérage et à l'analyse des syntagmes nominaux » (Bourigault et Jacquemin, 2000). Au lieu d'identifier à l'instar du logiciel Tropes, les relations du type « gouvernement→entreprise », l'algorithme mis au point par Bourigault est en mesure de distinguer les syntagmes nominaux, correspondant à des séquences de mots du type « gouvernement des entreprises » ou « gouvernement de l'entreprise »[12].

De tels syntagmes, identifiés dans le corpus textuel, s'élèvent au nombre de 978, et forment une ontologie inédite de la gouvernance d'entreprise. Celle-ci peut être organisée, par le biais d'un classement simple des syntagmes sur la base de leurs têtes communes, de telle sorte que pour un terme tel qu'« administrateurs », on observera différentes expansions possibles du type : « externes », « internes », « indépendants » ou « salariés ». Au delà, on peut envisager une catégorisation des têtes de syntagmes (voir tableau 1). Ce classement de deuxième niveau étant réalisé, une nouvelle présentation du cadre conceptuel de la gouvernance peut alors être proposée. Celui-ci est constitué de huit groupes principaux, à partir d'un total de 216 têtes correspondant à 863 syntagmes[13].

n°	Catégories de têtes de syntagmes	Nombre de têtes retenues	Exemples de têtes de syntagmes
1.	Acteurs	31	dirigeant, actionnaire, administrateur, …
2.	Jeu des acteurs	27	contrôle, confiance, enracinement, …
3.	Comportement et décisions du dirigeant	16	latitude, investissement, partage, …
4.	Cadre d'analyse qualitatif	69	mécanisme, processus, risque moral, …
5.	Cadre d'analyse quantitatif	33	indicateurs, mesures, performance, valeur, …
6.	Cadre institutionnel	7	firme, marché, réseau, …
7.	Cadre juridique	11	clauses, contrats, créances, droit, …
8.	Contextualisation du discours	22	contexte, degré, niveau, mode, secteur, type, …

Tableau 1. Têtes de syntagmes classées par catégories

[12] Dans ce cas, les deux syntagmes « gouvernement des entreprises » et « gouvernement de l'entreprise » ont pour tête commune le terme « gouvernement ».

[13] Sur la liste des 978 syntagmes initiaux, 115 ne sont pas classés dans les huit groupes proposés (cf. tableau 1), car ils ne représentent qu'un intérêt limité.

Ainsi classée, l'ontologie de la gouvernance dévoile le mode de structuration d'une théorie chargée de conceptualiser et de modéliser un jeu complexe d'acteurs. Replacés dans leurs cadres institutionnels et juridiques, et leur contexte spécifique, les problèmes de gouvernance sont alors appréhendés en s'appuyant sur des notions qualitatives et quantitatives, et leurs éventuelles interactions. À titre d'exemple, la rémunération du dirigeant appartient bien au champ du quantitatif. Mais il n'en demeure pas moins qu'elle renvoie aussi à la mesure de la performance, elle-même tributaire d'aspects qualitatifs comme par exemple les manipulations de l'information financière, ou l'existence d'asymétries d'information entre décideurs et investisseurs.

Au delà de cette analyse, il est possible d'approfondir la connaissance de ces différentes catégories en focalisant la recherche sur l'étude des mises en relations.

3.3. *Relations entre acteurs*

De toute évidence, la variété des acteurs de la gouvernance obtenue dans la catégorie 1 des têtes de syntagmes va bien au delà du triptyque fondamental dirigeants-actionnaires-conseil d'administration résultant du graphe 1. Afin de réunir l'ensemble le plus large possible des acteurs pris en considération dans le corpus textuel, la liste des « références utilisées » fournie par Tropes a été exploitée exhaustivement. Un premier « scénario » a pu ainsi être mis en œuvre grâce au logiciel Tropes, celui-ci retenant une série de 25 acteurs[14] tirés de la catégorie 1 et 13 autres obtenus à partir de la liste intégrale des références utilisées[15]. Compte tenu de cette liste, 3061 citations d'acteurs ont pu être référencées. Sept acteurs concentrent à eux seuls quelque 80% des 3061 citations. Il s'agit du dirigeant, de l'actionnaire, du conseil d'administration, des administrateurs, de la banque ou des banquiers, des créanciers et des 'stakeholders'. Cela confirme bien que le corpus analysé est centré sur un effectif restreint d'acteurs, même si l'on admet que le terme 'stakeholder' contribue à élargir le nombre des parties prenantes prises théoriquement en considération.

4. Conclusion

Cette recherche a montré que les cartographies d'acteurs et de concepts, obtenues à partir des analyses sémantique et syntaxique, peuvent fournir des représentations inédites, susceptibles d'aider à l'inventaire et à l'analyse de l'évolution des idées en gouvernance d'entreprise. Elles permettent de visualiser les éléments centraux du discours, ainsi que de mieux comprendre et d'anticiper les évolutions de ce champ de connaissances. De façon générale, les résultats obtenus invitent les chercheurs, désireux d'approfondir les sujets classiques de la gouvernance ou d'en explorer de nouveaux, à relire les textes fondateurs, tout en s'interrogeant sur les contextes d'utilisation des concepts et de mise en scène des acteurs. Ce travail pourra être poursuivi à partir à partir d'un corpus élargi de publications. Une telle ingénierie des connaissances, spécialisée en gouvernance d'entreprise, pourrait ainsi émerger, en améliorant sans

[14] Les acteurs tirés de la catégorie 1 sont les suivants (le nombre placé entre parenthèses indique le nombre de citations dans le corpus): dirigeant (1164), actionnaire (595), conseil d'administration (189), administrateurs (155), banque ou banquier (133), créancier (118), stakeholder (87), partenaire (68), client (61), investisseur (54), apporteur (44), fournisseur (38), propriétaire (37), pouvoirs publics (33), bureaucrate (31), parties prenantes (31), salarié (28), prêteur (24), consommateur (12), équipe dirigeante (8), hommes politiques (8), collectivités locales (5), analystes financiers (4),commission bancaire (3), comité d'audit (2).

[15] Les acteurs complémentaires obtenus à partir des « références utilisées » sont les suivants (le nombre placé entre parenthèses indique le nombre de citations dans le corpus): entrepreneur (36), groupe (26), décideur (14), détenteur (13), syndicat (10), avocat (7), économiste (6), chercheur (5), commission des opérations de bourse ou COB (3), législateur (3), experts-comptables (2), expert(indépendant) (2), et lobbies (2).

cesse la structuration sous-jacente des raisonnements, et en ouvrant par là même de nouveaux axes d'interrogations et d'approfondissements des connaissances.

Références

Allard-Poesi F., Drucker-Godard C. et Ehlinger S. (1999). Analyses de représentations et de discours. In chapitre 16 de Thiétart R.A. *et al*. (Eds), *Méthodes de recherche en management*. Dunod : 449-475.

Bourigault D. et Assadi H. (2000). Analyse syntaxique et analyse statistique pour la construction d'ontologie à partir de textes. In Charlet J., Zacklad M., Kassel G. et Bourigault D. (Eds), *Ingénierie des connaissances. Tendances actuelles et nouveaux défis*. Eyrolles : 243-255.

Bourigault D.et Jacquemin C. (2000). Construction de ressources terminologiques. In Pierrel J.-M. (Ed.), *Ingénierie des langues*. Hermès : 215-233.

Charreaux G. (Ed.) (1997). *Le gouvernement des entreprises – Corporate Governance – Théories et faits*. Economica.

Charreaux G. (2000b). Gouvernement d'entreprise et comptabilité. *Encyclopédie de Comptabilité, Contrôle de Gestion et Audit*, article (55) : 743-756.

Charreaux G. (2002). Variation sur le thème : « A la recherche de nouvelles fondations pour la finance et la gouvernance d'entreprise ». *Finance Contrôle Stratégie*, vol. (5/3), septembre: 5-68.

Cossette P. (1999). L'analyse de la pensée de F.W. Taylor à l'aide de la cartographie cognitive. Journées de l'AIMS, Ecole centrale de Paris.

Cossette P. et Audet M. (1994). Qu'est-ce qu'une carte cognitive ? In Cossette P. (sous la direction de), *Cartes cognitives et organisations*. Editions Eska, Les Presses de l'Université de Laval : 13-33.

Freeman R.E. (1984). *Strategic Management : A Stakeholder Approach*. Pitman.

Ghiglione R. et Blanchet A. (1991). *Analyse de contenu et contenus d'analyse*. Dunod.

Ghiglione R., Landré A., Bromberg M. et Molette P. (1998). *L'analyse automatique des contenus*, Dunod.

La lemmatisation de l'arabe non classique

Laurence Tuerlinckx

U.C.L. – Institut orientaliste – Centre d'Études sur Grégoire de Nazianze
Place Blaise Pascal 1 – 1348 Louvain-la-Neuve – Belgique
tuerlinckx@ori.ucl.ac.be

Abstract

The current tools for the lemmatization and automatic analysis of the Arab language were conceived for classical or standard Arabic (literary corpora, newspapers...). Philological research on ancient texts written in non classical language (Christian texts in Middle Arabic in our case) requires a system which takes into account the graphic and syntactic variations of this language level. A program of formal lemmatization, adapted to these characteristics, was elaborate with the aim of creating lemmatized concordances and indexes. This software is based not on automatic analysis, but on recognition of forms already encountered, which makes it possible to gather the various forms under the same classical lemma, whatever their actualization in the texts. The lemmata are tagged with regard to their morpho-syntactic class as defined by Arab grammarians, and are classified by root, according to the model of traditional dictionaries. The philological options and methodological thoughts behind the treatment of this language (characterized by agglutination and non-vocalization) will be presented, as well as some prospects about exploiting this new data base.

Résumé

Les outils actuels de lemmatisation et d'analyse automatique de la langue arabe ont été conçus pour l'arabe classique ou standard (corpus littéraires, journaux…). Les recherches philologiques portant sur des textes anciens rédigés en une langue non classique (textes chrétiens en moyen arabe dans notre cas) nécessitent donc un système qui prenne en compte les variations graphiques et syntaxiques de cet état de langue. Un programme de lemmatisation formelle, adapté à ces particularités, a été élaboré dans le but premier de créer des concordances et index lemmatisés. Ce logiciel, basé non sur l'analyse automatique, mais sur la reconnaissance des formes déjà rencontrées, permet de classer les formes sous un même lemme classique, quelle que soit leur actualisation dans les textes. Les lemmes sont étiquetés relativement à leur classe morphosyntaxique telle que définie par les grammairiens arabes et sont regroupés par racine, suivant le modèle des dictionnaires classiques. Les options philologiques et la réflexion méthodologique concernant le traitement de cette langue (caractérisée par l'agglutination et la non vocalisation) seront présentées, ainsi que quelques perspectives d'exploitation de cette nouvelle base de données.

Mots-clés : arabe, lemmatisation, forme, lemme, racine, analyse, morphologie, philologie

1. Introduction

Depuis plusieurs décennies déjà, des recherches sont poursuivies dans le cadre du traitement automatique de la langue arabe. L'un des premiers théoriciens de ce domaine, David Cohen propose un essai d'analyse automatique dès 1961 (Cohen, 1961/1970). Partant d'une analyse morphologique minimaliste, basée sur le principe que toute forme linguistique arabe se traduit en schème et racine, les recherches vont se développer depuis les premiers travaux sur le lexique et la morphologie jusqu'à la mise au point d'analyseurs automatiques, de systèmes d'indexation, de correcteurs, etc. De nombreux projets sont en cours et il existe des bases de données disponibles proposant des corpus divers sous forme électronique, traités automati-

quement[1]. Un traitement automatique de l'arabe est donc non seulement possible, mais réalisé et en cours de perfectionnement (notamment dans le domaine de la traduction automatique).

Cependant, dans le cadre bien précis de nos recherches, nous avons été amenés à créer notre propre programme de lemmatisation. Après un exposé du problème, nous présenterons ce logiciel, ainsi que la réflexion méthodologique qui a présidé à son élaboration. Les premiers résultats obtenus et surtout les développements envisagés seront ensuite décrits.

2. La lemmatisation des textes en arabe non classique

Les outils disponibles pour le traitement automatique de l'arabe, conçus pour la langue classique, ne sont pas adaptés aux recherches philologiques basées sur des textes anciens rédigés en arabe non classique.

2.1. *Problème des états de langue en arabe*

Dès les premières réflexions sur le traitement automatique de l'arabe, le champ d'investigation a été défini : « *Par ARABE il est entendu ici L'ARABE LITTÉRAIRE, langue écrite de la littérature et de la presse, et parlée ordinairement à la radio, dans les cours et conférences universitaires et scientifiques, dans les discours officiels, etc., dans tous les pays arabes. Il devra être également admis, bien que cela n'ait aucune incidence sur la théorie, qu'il s'agit de L'ARABE MODERNE, ce qui permettra dans les essais éventuels de négliger des faits qui peuvent caractériser l'usage de la littérature la plus ancienne.* » (Cohen, 1970 : 49)[2]. Cette limitation du domaine pris en compte, justifiée du point de vue méthodologique, est liée à la question particulière des divers états de langue. L'arabe, langue sacrée du Coran, connaît une grande stabilité dans un créneau bien précis qui est celui de la littérature classique et des milieux de l'enseignement, et aujourd'hui, le domaine de la culture officielle, de la presse, etc. C'est l'arabe standard ou littéraire, universellement partagé par les lettrés de tous les pays arabes. Par contre, parallèlement à cette lignée relativement figée si ce n'est l'élargissement du lexique, il existe de nombreuses branches s'écartant plus ou moins de la norme. L'arabe dialectal dans toutes ses variétés, essentiellement oral, et le moyen arabe (état intermédiaire entre le dialectal et le classique) essentiellement écrit, sont autant de réalisations différentes d'une même source. Suffisamment proches pour constituer une seule et même langue, suffisamment éloignées pour ne pas s'intégrer dans les systèmes de traitement automatique, ces variétés linguistiques demandent une autre approche.

2.2. *La lemmatisation de l'arabe*

Sur le plan théorique, les principes généraux de la lemmatisation s'appliquent également à la langue arabe et les définitions établies dans le cadre du « Projet de recherche en lexicologie grecque », nous ont servi de postulat de départ : « *Au niveau du lexique de la langue, les* lemmes *servent d'adresses lexicales aux* formes*; les* formes *sont, au niveau du discours, les actualisations des* lemmes*. Classer les* formes *sous les* lemmes *adéquats, et définir les* lemmes *de manière pertinente, contraint le "lemmatiseur" décrivant le vocabulaire d'un auteur*

[1] Par exemple: The An-Nahar Newspaper Text Corpus, proposé sur le site de l'ELDA (http://www.elda.fr) ou l'analyseur morphologique mis au point par Ken Beesley : the Xerox Arabic Morphological Analysis and Generation, disponible en version demo sur le site du Xerox Research Centre Europe : http://www.xrce.xerox.com/competencies/content-analysis/arabic/ Le dictionnaire électronique Sakhr : http://www.aramedia.com/diction.htm. Nous ne savons pas ce que sont devenus le projet DIINAR-MBC et son analyseur AraParse (Dichy, 2001 : 1 ; Ouersighni, 2001 : 1).

[2] Cette limite du champ d'investigation des recherches est unanimement partagée par tous ceux qui travaillent sur le traitement automatique de l'arabe. Voir par exemple les titres des conférences dans ACL/EACL, 2001.

et (re)construisant progressivement le lexique d'une langue, à porter un regard sur cette langue et à l'analyser »[3]. Si les principes développés en fonction du grec ancien valent pour l'arabe, il y a cependant certaines particularités de cette langue qui demandent une adaptation dans la manière de concevoir la lemmatisation.

2.2.1. *La lemmatisation de l'arabe standard ou classique*

Outre l'orientation droite-gauche, les principales caractéristiques de la langue arabe, langue sémitique, par opposition aux langues indo-européennes sont l'agglutination, la structure particulière combinant schème et radical, et la non vocalisation.

Le *mot* ou l'*unité graphique*, « *suite de graphèmes entre deux blancs* » (ou signes de ponctuation) correspond le plus souvent en arabe non pas à une *forme* ou « *unité susceptible de figurer sous une entrée lexicale ou lemme* »[4], mais à une suite de formes collées les unes aux autres. Les mots du texte sont des *formes agglutinées* (ex. conjonction + préposition + article + nom) ; théoriquement cela nécessite une étape préliminaire de séparation des formes, mais concrètement, c'est la forme graphique entière qui est lemmatisée, c'est-à-dire ramenée à une succession de lemmes distincts.

La structure du mot arabe est donc décomposable en cinq éléments : proclitique, préfixe, base, suffixe et enclitique. La *base* est une combinaison de lettres radicales (le plus souvent trois) et d'un schème (redoublement, augment, voyelles longues et brèves, etc.). La base – avec préfixe et suffixe – forme le *noyau lexical*, éventuellement entouré d'extensions (Dichy, 1997 : 296-297 et 2001 : 3-4). Les analyseurs automatiques travaillent à partir des combinaisons valides de ces éléments.

Les voyelles brèves ne sont pas indiquées dans les textes courants. En règle générale, seuls le Coran et les textes à vocation didactique sont vocalisés. Cette caractéristique entraîne un haut degré d'ambiguïté, les formes non vocalisées peuvent représenter un nombre élevé d'actualisations différentes. Le traitement automatique de l'arabe doit pouvoir traiter des textes vocalisés comme des textes non vocalisés.

2.2.2. *La lemmatisation de l'arabe non classique : l'exemple du projet « Grégoire de Nazianze »*

Le projet « Grégoire de Nazianze » a pour objet l'ensemble des versions orientales d'une même œuvre : les discours de Grégoire de Nazianze, Père de l'Église et auteur grec byzantin (4e siècle)[5]. La version arabe de ces discours, datable du 10e siècle, témoigne de cet état de langue appelé « moyen arabe » (voir 2.1.) ; les éditions (Grand'Henry, 1988 et 1996 ; Tuerlinckx, 2000) ne corrigent pas le texte, ni ne le standardisent. Le texte édité, à lemmatiser ne s'inscrit donc pas dans le registre classique. Les textes rédigés en moyen arabe présentent des particularités graphiques, morphologiques et syntaxiques qui s'écartent de l'arabe classique (Blau, 1966). Les modifications graphiques ont pour conséquence immédiate que l'analyseur conçu pour l'arabe classique ne reconnaît pas la forme entrée. Certaines lettres faibles faisant partie de la racine (', *w*, *y*) reçoivent un traitement différent en moyen arabe (particulièrement dans la conjugaison) ce qui empêche la reconnaissance des formes. L'ordre

[3] Ce projet est développé à l'Institut orientaliste de l'Université catholique de Louvain, à Louvain-la-Neuve. Cette citation est extraite de la présentation du projet et du *Thesaurus Patrum Graecorum* sur le site http ://tpg.fltr.ucl.ac.be. Voir aussi Coulie, 2003.

[4] Définitions de P. Tombeur, citées dans Coulie (1996 : 37).

5 Site du projet : http://nazianzos.fltr.ucl.ac.be

des mots fixé par des règles précises en arabe classique, n'est pas respecté et une simplification de la syntaxe (disparition des modes, modification des genres, etc.) rend invalides certaines grammaires sous-jacentes à l'analyse automatique et à la levée d'ambiguïté. De plus le lexique particulier des textes chrétiens comporte un nombre important et illimité de termes d'origine non arabe, ne rentrant donc pas dans la structure de base vue plus haut[6].

Jusqu'à présent, les éditions de cette version arabe ne comportent pas d'index, car seul un index lemmatisé permet des recherches sur le lexique et la langue. Un index des formes du texte est facilement réalisable, mais sans intérêt étant donné le nombre important d'éléments adventices placés devant la base.

L'approche multilingue du projet vise à la comparaison des versions latine, arménienne, syriaque, arabe et géorgienne ; cette option nécessite un même système de référence au texte grec (langue source) dans toutes les langues traitées. Le système de référence doit être compatible avec celui de la concordance lemmatisée de ces textes dans le *T.P.G.* (Mossay et Cetedoc, 1990), en vue de la réalisation de lexiques multilingues et d'alignement des contextes.

3. Le programme de lemmatisation « GLOR.ARABE »

Afin de répondre aux besoins spécifiques mentionnés ci-dessus, un logiciel de lemmatisation a été créé grâce aux efforts conjoints d'arabisants et d'informaticiens[7].

Ce logiciel ne s'appuie pas sur des traitements automatiques de l'arabe développés ailleurs, et ceci en raison d'un choix, motivé d'une part par les exigences précises que nous posons en tant que philologues, d'autre part par les conclusions négatives obtenues à l'issue d'une enquête concernant l'accessibilité aux outils informatiques existants (contacts restés sans réponse, manque de suivi dans les projets exposés, incompatibilité avec le projet multilingue de notre équipe et surtout champs et objectifs trop éloignés des nôtres).

3.1. *Description de la méthode et des caractéristiques techniques*

Le programme GLOR.ARABE résulte de la mise au point de deux applications : l'une concerne les prétraitements, l'autre consiste en un outil interactif d'analyse linguistique élémentaire. Le langage de programmation utilisé est *MS Visual Basic* et les procédures et formulaires sont réalisés sous *MS Access*.

3.1.1. *Prétraitements*

Les opérations de prétraitement comportent quatre points :

– la découpe du texte en unités graphiques ou « formes agglutinées », accompagnées d'une série de références (auteur, œuvre, livre, chapitre, paragraphe, ligne, nœud ou numéro d'ordre du mot) ;

– la réorganisation du texte en vue d'une version finale (après ajouts, suppressions de mots, corrections, modification des sauts de lignes) ;

[6] Le même phénomène existe bien sûr en arabe classique ; Dichy appelle « pro-bases » ces noms issus d'emprunt. (Dichy, 1997 : 296).

[7] Les principaux concepteurs de ce logiciel sont Boris Maroutaeff, informaticien de l'équipe Informatique Facultaire, et Aurélie Gribomont, doctorante à l'Institut orientaliste ; les principes théoriques ont été discutés et établis par J. Grand'Henry et L. Tuerlinckx (tous sont membres de la Faculté de Philosophie et Lettres de l'Université catholique de Louvain à Louvain-la-Neuve). Ce logiciel a été baptisé GLOR.ARABE du nom du Département d'Études Grecques, Latines et Orientales de cette faculté.

– la réalisation d'unités de contextes ou concordances (5 mots avant le mot-clé, 5 mots après) ;

– l'établissement d'index et de tables en vue de la publication d'un index lemmatisé des textes.

3.1.2. *Analyse linguistique*

L'analyse linguistique développe trois procédures :

– l'encodage d'entrées de dictionnaires ou « lemmes » pour les différentes formes du texte (création d'un *Thesaurus* arabe) ;

– la lemmatisation (première) automatique et interactive au fur et à mesure de l'enrichissement du *Thesaurus* des lemmes (reconnaissance de formes et de lemmes) ;

– une procédure de navigation à l'intérieur du texte et du *Thesaurus* en vue de la proposition d'analyse.

3.1.3. *Caractéristiques techniques*

Ces applications requièrent le système d'exploitation *MS Windows XP Pro* et le traitement de texte *MS Office XP Pro*. Les données textuelles sont du format Unicode (police Arial unicode MS arabe) ; contrairement à ce qui se fait habituellement en traitement automatique de l'arabe, nous travaillons directement sur caractères arabes et non sur transcription.

Le texte peut provenir d'un fichier *Mac* (*Nisus* par exemple) ; il doit alors être converti en fichier *Windows* à l'aide du programme *Mac2Win*.

3.2. *Le formulaire de lemmatisation*

Le texte encodé doit être nettoyé de toute ponctuation, lignes blanches, etc., les sauts de ligne doivent être marqués (dans le cas d'un texte déjà édité, il faut respecter les lignes de l'édition). Le texte ainsi préparé est converti en un fichier '.txt' qui est rattaché à la base de donnée sous un numéro d'identification. La lemmatisation peut alors commencer. Après sélection de l'œuvre à traiter dans la liste des textes encodés, un formulaire de lemmatisation s'affiche, avec une série de champs remplis (contexte, références, cadre de propositions de lemmes) ou à remplir (cadre de lemmatisation).

La ligne supérieure situe le terme à lemmatiser dans son *contexte*. La *forme agglutinée* est le « mot graphique » composé de une ou plusieurs formes. Les références sont celles de la division interne à l'œuvre, le *nœud* étant le numéro d'ordre du mot graphique dans le texte. Ce numéro est définitivement attribué au mot ; les additions ou suppressions de mots suite à une correction d'éditeur par exemple, se traitent par le biais du *nœud caché*. Ce système de référence est parfaitement compatible avec le système de référence utilisé par les lemmatiseurs du *T.P.G*. Le cadre *lemmatisation* comportera autant de lignes qu'il y a de formes dans la forme agglutinée, chaque forme reçoit un *lemme*, auquel sont attachées une *nature* et une *racine*. Le *numéro de lemme* indique la position de la forme en question à l'intérieur de la forme agglutinée ; ce numéro n'est donc pas lié au lemme lui-même, mais à sa forme actualisée. Enfin, la case *emprunt* est cochée lorsque le lemme n'est pas un mot arabe, mais une transcription ou une adaptation d'un mot étranger. Le cadre *propositions* fournit, le cas échéant, des propositions de lemme pour la ou les différentes formes composant la forme agglutinée.

3.3. *Principe de la reconnaissance des formes*

Le principe de base de ce programme est non pas l'analyse, mais la reconnaissance de formes déjà rencontrées, aussi aberrantes et peu classique soient-elles. La reconnaissance fonctionne à deux niveaux différents : au niveau de la forme agglutinée et au niveau du lemme de chaque forme séparée.

Face à une forme du texte, le logiciel propose les différentes combinaisons de lemmes ou les différents lemmes qui ont déjà été attribués à cette même forme dans un quelconque texte lemmatisé. Il suffit alors de choisir le(s) lemme(s) qui convien(nen)t et de le(s) rattacher au mot traité.

Le second niveau de reconnaissance, intervient dans la lemmatisation de la ou des forme(s) d'un mot graphique. Lorsqu'on commence à encoder le lemme d'une forme non reconnue, un menu déroulant propose les lemmes déjà enregistrés et une recherche peut se faire sur le lemme entier ou sur les premières lettres (recherche partielle). Le résultat fournit un ou plusieurs lemme(s), accompagné(s) de leur nature, de leur racine, et de un ou plusieurs numéro(s) d'ordre. (Ex. l'article proposé en 1re, 2e ou 3e position).

Si aucun lemme proposé ne convient, il faut encoder un nouveau lemme et lui adjoindre ses éléments distinctifs (numéro, nature, racine, emprunt). L'actualisation de ces données entraîne la reconnaissance du nouveau lemme lorsqu'une forme identique sera rencontrée.

4. Aspects théoriques de la lemmatisation arabe

Les champs à remplir dans le formulaire de lemmatisation sont autant d'éléments à déterminer qui, ensemble, définissent le lemme. Chacune de ces déterminations s'appuie sur des choix théoriques et pose des problèmes spécifiques à résoudre.

4.1. *Lemme – Nature – Racine : définitions*

La lemmatisation réduit les formes à leur entrée de dictionnaire ; en arabe comme dans les autres langues, on se conforme à l'usage des dictionnaires classiques[8].

4.1.1. *Lemme*

Le lemme est l'entrée de dictionnaire, il s'agit donc de la forme classique et entièrement vocalisée à laquelle se rattache la forme du texte (classique ou non, vocalisée ou non) :

– les *verbes* sont ramenés à la 3e pers. sg. de l'accompli actif, sauf dans le cas de certains verbes figés n'ayant qu'une conjugaison partielle ;

– les *noms variables* sont ramenés à la forme du nominatif sg. (masc. pour les noms qualificatifs), les *noms invariables* à leur forme classique vocalisée (masc. sg. pour les pronoms) ;

– les *particules* sont ramenées à leur forme classique vocalisée.

À la différence d'autres langues comme le grec, on ne peut dire qu'il y a en arabe des « *formes indéclinables et indéclinées dans un texte, qui sont donc leur propre "forme de départ" ou "entrée de dictionnaire"* » (Coulie, 1996 : 38). La non vocalisation et la variation orthographique de la langue (voir 4.2.2.) demandent une lemmatisation distincte de la forme.

Les lemmes composés sont gardés comme composés si le composé lui-même fait l'objet d'une entrée dans les dictionnaires classiques. La nature qui lui est attribuée est celle du com-

[8] Notre principal dictionnaire de référence est celui de la Ligue arabe : ALECSO (1989).

posé, la racine retenue est celle du premier élément, afin d'assurer au composé sa juste place dans l'ordre alphabétique.

4.1.2. *Nature*

À chaque lemme est associée une étiquette précisant sa nature morphosyntaxique. En arabe, les catégories de mots se réduisent à trois : les verbes, les noms et les particules. Certains grammairiens ajoutent une catégorie « instrument » recoupant plus ou moins celle des particules ; nous avons limité cette catégorie à l'article, afin de le distinguer des autres particules. L'analyse suivie ici est celle des grammairiens arabes[9], non celle des arabisants occidentaux, auteurs de grammaires anglicisées ou francisées. Ainsi, par exemple, il n'y a pas de catégorie « adverbe » ou « préposition » et les participes ne sont pas rattachés à la catégorie « verbe », mais à la catégorie « nom ». Ces catégories se subdivisent en notions plus précises : 85 natures différentes ont été distinguées.

En dehors de l'analyse automatique, cet étiquetage morphosyntaxique est utile pour plusieurs raisons :

– il permet de lever des ambiguïtés entre certains lemmes et de distinguer certains lemmes homographes ;

– il intervient dans la structure du dictionnaire, puisque sous les racines, les lemmes sont classés selon une préséance due à leur nature et non selon l'ordre alphabétique ;

– il fournit un élément intéressant dans l'étude du lexique d'un texte.

4.1.3. *Racine*

Le plus souvent trilitère, la racine est la suite des consonnes formant le radical du mot ; à chaque racine correspond un champ sémantique. La racine est un élément important dans les langues sémitiques : associée à un schème, elle forme la base du mot (voir 2.2.1.). Les lemmes sont tous rattachés à une racine ; ce sont les racines qui sont classées alphabétiquement dans le dictionnaire, non les lemmes (voir 4.1.2.).

Racine est à prendre ici au sens large : il s'agit, sur le plan pratique, de fournir une référence qui détermine la place du lemme dans le dictionnaire. Dans certains cas, sous « racine » se trouve une suite de lettres qui ne constitue pas un radical, mais reprend le lemme dépourvu de ses voyelles brèves et d'éventuels éléments adventices : ce sont les *lemmes-racines* (noms propres, emprunts, transcriptions, onomatopées, particules, etc.). La case « emprunt » cochée est souvent associée à un lemme-racine.

La racine permet également de distinguer des lemmes homographes : deux combinaisons de lettres radicales et de schèmes différents peuvent aboutir au même résultat, alors qu'il s'agit de deux lemmes distincts.

4.2. *Problèmes particuliers*

4.2.1. *Homographie et spécifications*

Des spécifications sont ajoutées aux lemmes, non seulement pour fournir une information complète sur les entrées du dictionnaire, mais aussi pour réduire les cas d'homographie. Ces spécifications consistent à ajouter

[9] La grammaire suivie est en grande partie celle de Chartouni (Grand'Henry, 2000).

– à tous les verbes de la 1re forme: la forme conjuguée à l'inaccompli. Par contre, la voyelle de l'accompli est précisée dans la nature (V1a, V1i, V1u), parce que cette voyelle régit l'ordre des verbes 1 dans le dictionnaire. Des voyelles de conjugaison différentes peuvent s'accompagner de sens distincts et donc justifier des lemmes distincts.

– à certains noms singuliers: le(s) pluriel(s) brisé(s) (irréguliers). Des pluriels différents peuvent prendre des significations nettement distinctes et dans ce cas, ils justifient la création de lemmes distincts.

– à certaines particules : une notification précisant le type de particule (notification ajoutée à la nature) ou son usage (notification ajoutée au lemme). Des usages nettement différents d'une même particule justifient la création de lemmes distincts.

4.2.2. *Variations orthographiques*

Les variations orthographiques sont répandues, même à l'intérieur d'un même texte. L'éditeur d'un texte ancien peut rendre dans son édition la variété présente dans les manuscrits, si ses principes d'édition le justifient. Cela ne pose pas de problème pour la lemmatisation, puisque le lemme suit toujours l'orthographe classique. Cependant une variation peut subsister d'un dictionnaire à l'autre, particulièrement au niveau de la vocalisation ; en ce cas, les différentes vocalisations sont données sous un même lemme. Le dédoublement du lemme ne se justifie que si la variation vocalique s'accompagne d'une modification de sens.

4.2.3. *Ambiguïté*

L'absence de voyelles brèves dans les textes et l'écriture sporadique de signes tels que le redoublement ou l'allongement, entraînent une très forte ambiguïté. Généralement, cette ambiguïté est facilement résolue par l'association de formes, le sens, le contexte, etc. Dans certains cas, peu fréquents, rien ne justifie le choix d'un lemme plutôt qu'un autre, les différents lemmes possibles sont alors mentionnés, mais avec le même numéro d'ordre dans la forme, indiquant ainsi qu'ils occupent la même position.

4.2.4. *Référence*

Les noms propres et leurs dérivés, ainsi que les emprunts à d'autres langues posent un problème de référence. Ces formes présentent une grande variété orthographique dans les textes et il n'existe dans le domaine arabe, ni ouvrage de référence absolue, ni bases de données suffisamment développées pour pouvoir utiliser un critère tel que la fréquence d'un usage ou son ancienneté, comme cela peut se faire en grec par exemple.

5. Résultats obtenus et développements

Le travail de lemmatisation n'en est qu'à ses débuts, mais un dictionnaire commence à se former et les requêtes permettent d'effectuer certaines recherches sur le lexique des textes étudiés.

5.1. *Résultats – État d'avancement*

Deux discours de Grégoire de Nazianze ont été lemmatisés (Discours 1 et 27), deux autres sont en cours de lemmatisation (Discours 45 et 40) ; 24684 formes séparées ont reçu un lemme, ce qui correspond à 2692 lemmes différents encodés.

En quelques chiffres[10] :

	Discours 1	***Discours 27***	***Discours 45***	***Discours 40***
Formes agglutinées	825	2173	(2650 / 6865)	(7250 / 11181)
Formes séparées	1297	3407	(3988)	(15121)
Verbes	152 (11,7%)	378 (11%)	427 (10,7%)	1888 (12,4%)
Noms	611 (47,1%)	1558 (45,7%)	1858 (46,5%)	6679 (44,1%)
Particules	371 (28,6%)	1001 (29,3%)	1222 (30,6%)	4818 (31,8%)
Instruments/ articles	160 (12,3%)	467 (13,7%)	479 (12%)	1738 (11,4%)

On constate la stabilité des rapports entre les différentes catégories de lemmes, les pourcentages sont proches d'un texte à l'autre. Le rapport entre formes agglutinées et formes séparées varie peu : en moyenne chaque forme du texte se compose de 1,5 forme. (Plus dans le Discours 40, mais la lemmatisation en cours est ici sélective et ne reflète pas la moyenne).

Les requêtes et états en *Access* permettront de fournir un index lemmatisé des textes édités et en cours d'édition, et de mener dès à présent certaines recherches approfondies sur le lexique et les caractéristiques de cette version arabe. (Par ex. la fréquence élevée de telle catégorie morphosyntaxique comme résultat d'un procédé de traduction bien précis).

5.2. *Développements*

Cette nouvelle base de donnée demande à être enrichie de textes arabes issus d'époques et de milieux divers, musulmans et chrétiens, classiques et non classiques. Cela permettrait de développer des études synchroniques et diachroniques du lexique arabe.

Nous avons grandement bénéficié de l'expérience des auteurs du *T.P.G.* (*Thesaurus Patrum Graecorum*) et du *D.A.G.* (Dictionnaire automatique grec)[11] ; ces travaux nous ont servi de modèles et ont nourri notre réflexion sur la lemmatisation. La compatibilité de notre *D.A.A.* (Dictionnaire automatique arabe) naissant, avec le *D.A.G.* rend possible des développements multilingues. Le lien de traduction unissant une ou plusieurs forme(s) du texte grec à une ou plusieurs forme(s) du texte arabe, sera automatiquement transformé en un lien de lemme(s) à lemme(s), et le système commun de référence au texte facilitera la correspondance entre le texte traduit et sa source. Les développements récents du *D.A.G.* en un *D.D.G.* (Dictionnaire dérivationnel du grec) qui permet une interrogation à partir des morphèmes grecs (Gérard et Kindt, 2004) sont particulièrement intéressants pour nos recherches : le traducteur arabe avait tendance à décomposer le mot grec en morphèmes de la même manière que le fait le *D.D.G.*, avant de le faire passer dans la langue arabe, souvent élément par élément. Le lexique bilingue de nos textes devra se construire sur le lien de morphème(s) à lemme(s) plus que de lemme(s) à lemme(s) et l'exploitation du *D.D.G.* s'avère dès à présent une aide précieuse pour la compréhension des techniques de traduction (du grec à l'arabe) mises en œuvre dans les textes que nous étudions.

Le logiciel GLOR.ARABE lui-même est appelé à de nouveaux développements, car les syriacisants de notre équipe de recherche (CEGN) sont intéressés par une lemmatisation de ce

[10] État de la base de données en décembre 2003.

[11] Présentation du *D.A.G.* et de ses développements sur le site du *T.P.G.* : http://tpg.fltr.ucl.ac.be/PAGE.HTM.

type. L'adaptation du programme au syriaque est en cours d'élaboration ; le syriaque est une langue proche de l'arabe, appartenant au même groupe des langues sémitiques et partageant donc un certain nombre de caractéristiques communes. Cependant, des modifications sont nécessaires notamment en raison de l'imperfection du tri alphabétique syriaque en *Access*. De plus, certaines options théoriques sont différentes en ce qui concerne l'analyse morphologique ou les spécifications à ajouter aux lemmes.

Ces développements constituent un pas de plus dans la réalisation des projets multilingues de notre équipe.

Références

ACL/EACL (2001) *Workshop, Arabic Language processing: status and Prospects.* Toulouse 6 juillet 2001, http://www.elsnet.org/acl2001-arabic.html.

ALECSO (1989). *Al-mu'jam al-'arabiyy al-'asâsiyy*. ALECSO – Larousse.

Blau J. (1966). *A Grammar of Christian Arabic. Based mainly on south Palestinian texts from the first millennium. Corpus Scriptorum Christianorum Orientalium*, vol. (267 : Subsidia 27).

Grand'Henry J. (2000). *Grammaire arabe à l'usage des Arabes. Traduction française et commentaires des Éléments d'arabe, morphologie et syntaxe, II de Rachid Chartouni (Beyrouth).* Série Pédagogique de l'Institut Linguistique de Louvain, vol. (24). Peeters.

Cohen D. (1961/1970). Cohen D (1961). Essai d'une analyse automatique de l'arabe. *La Traduction Automatique*, vol. (2/3) : 48-71. Republié dans Cohen D. (1970). *Études de linguistique sémitique et arabe.* Mouton : 49-78. Les citations renvoient aux pages de cette réédition.

Coulie B. (1996). La lemmatisation des textes grec et byzantins : une approche particulière de la langue et des auteurs. *Byzantion*, vol. (66) : 35-45. Article disponible on-line à partir de la page : http://tpg.fltr.ucl.ac.be/PAGE.HTM

Coulie B. (2003). Corpus Christianorum. Thesaurus Patrum Graecorum. In Leemans J. (Ed.), *Corpus Christianorum 1953-2003. Xenium Natalicum. Fifty Years of Scholarly Editing*. Brepols : 169-172.

Dichy J. (1997). Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. *Meta*, vol. (XLII, 2) : 291-306.

Dichy J. (2001). On lemmatization in arabic. A formal definition of the Arabic entries of multilingual lexical databases. In *Proceedings of ACL/EACL*.

Gérard R. et Kindt B. (2004). D'un dictionnaire de lemmatisation (*D.A.G.*) à un dictionnaire dérivationnel du grec ancien (*D.D.G.*). In *Actes des JADT 2004*.

Grand'Henry J. (1988). La version arabe du Discours 24 de Grégoire de Nazianze. Édition critique, commentaires et traduction. In Coulie B. (Ed.), *Versiones Orientales, repertorium Ibericum et studia ad editiones curandas* (*Corpus Christianorum. Series Graeca,* 20. *Corpus Nazianzenum,* 1). Brepols : 197-291.

Grand'Henry J. (1996). *Sancti Gregorii Nazianzeni opera. Versio Arabica antiqua, I. Oratio XXI* (*Corpus Christianorum. Series Graeca,* 34. *Corpus Nazianzenum,* 4). Brepols.

Mossay J. et Cetedoc (1990). *Thesaurus Sancti Gregorii Nazianzeni. Vol. 1. Orationes, Epistulae, Testamentum* (*Corpus Christianorum. Thesaurus Patrum Graecorum)*. Brepols.

Ouersighni R. (2001). A major offshoot of the DIINAR-MBC project: *AraParse*, a morpho-syntactic analyzer for unvowelled Arabic texts. In *Proceedings of ACL/EACL*.

Tuerlinckx L. (2000). *Sancti Gregorii Nazianzeni opera. Versio Arabica antiqua, II. Orationes I, XLV, XLIV* (*Corpus Christianorum. Series Graeca,* 43. *Corpus Nazianzenum,* 10). Brepols.

Quantifying semantic effects. The impact of lexical collocations on the inflectional variation of Dutch attributive adjectives

Jose Tummers, Dirk Speelman, Dirk Geeraerts

KU Leuven – RU Quantitative Lexicology and Variational Linguistics
Blijde Inkomststraat 21 – 3000 Leuven – Belgium
{jose.tummers;dirk.speelman;dirk.geeraerts}@arts.kuleuven.ac.be

Abstract

This paper wants to verify to what extent semantic effects contribute to the alternation between inflected and uninflected adjectives in a definite NP with a singular neuter head noun as observed in the *Corpus Gesproken Nederlands*. Existing accounts invoke the non-compositional semantics of the adjective-noun combination to interpret this case of inflectional variation, generally using highly idiomatic examples. We will argue for a combinatorial operationalisation of the semantic effect, based on lexical collocations. The influence of lexical collocations on the inflectional alternation will be measured by means of the log-likelihood test. Finally, the actual impact of lexical collocations on the inflectional alternation will be quantified integrating the semantic parameter together with other potential explanatory factors in a logistic regression analysis.

Keywords: variational linguistics, adjectival inflection, Dutch, lexical collocation, log-likelihood ratio, logistic regression.

1. Introduction

Dutch attributive adjectives are always inflected, taking an *-e* suffix (*het mooi-e paard* 'the beautiful horse'), except in indefinite NPs with a singular neuter head noun, where the adjective remains uninflected (*een mooi-∅*[1] *paard* 'a beautiful horse').

A special case within this inflectional system are definite NPs with a singular neuter head noun, which allow for an uninflected adjective next to the 'regular' inflected alternative, such as *het bijvoeglijk-∅ naamwoord* (lit. 'the attributive noun', hence 'the adjective') and *het koninklijk-∅ besluit* ('the royal decree'). Although a number of parameters influence this alternation, the most quoted explanation regards the semantics of the adjective-noun combination: whenever the adjective-noun combination has a non-compositional semantics, the adjective is said to remain uninflected. The present contribution aims at implementing this semantic parameter in order to test its empirical validity.

This paper is further structured as follows. Section 2 discusses the variation observed in definite NPs with a singular neuter head, demonstrating the need for an empirical analysis. In section 3, we propose to make the semantic parameter operational using the log-likelihood test to measure the lexical attraction between the adjective and the noun. Next, the results of the empirical analysis will be presented and evaluated by integrating the semantic parameter

[1] We use '∅' to identify the uninflected adjective. We do not take any position with respect to the existence of a zero-morpheme. This discussion would amply exceed the scope of this paper.

in a multifactorial analysis. We will conclude by summarising the most important findings and raise some issues for further research.

2. Inflectional variation in definite NPs with a singular neuter head noun

Dutch adjectival inflection challenges the classical model of agreement, where features present in the agreement domain, i.e. the NP, spell out the inflectional suffix, viz. *-e* or -∅. Definite NPs with a singular neuter head noun allow the use of both suffixes, resulting in an intricate case of variation. Several studies show that this inflectional alternation is governed by a series of parameters pertaining to different linguistic domains such as phonology, morphosyntax and lexical semantics as well as lectal factors such as register and the distinction between the two national varieties of Dutch (Lebrun and Schurmans-Swillen, 1966; de Rooij, 1980; Booij, 1992; ANS, 1997). We will briefly discuss the effects of the relevant parameters.

First, in NPs with a possessive pronoun as a determiner, such as *mijn sterk-∅ paard* ('my strong horse'), the uninflected adjective is more widespread than in sequences where the definite article *het* ('the') is used, such as *het sterk-∅ paard* ('the strong horse'). Second, rhythmic properties of the adjective may influence the inflection. The longer the adjective, the more natural the uninflected form seems to be, as in *het ongedifferentieerd-∅ corpus* ('the undifferentiated corpus'). This tendency is reinforced when the adjective ends in an unstressed or unvoiced syllable, such as *het ontzagwekkend-∅ paard* ('the awe-inspiring horse'). Third, semantically speaking, the use of the uninflected adjective can be observed when the adjective-noun sequence constitutes a semantic unity, such as *het Algemeen-∅ Nederlands* ('the standard variety of Dutch'). Fourth, from a lectal point of view, stylistic and geographical parameters modify the adjectival inflection. Stylistically, the uninflected alternative is considered to be more widespread in less formal contexts. Geographically, Belgian Dutch favours the uninflected adjective compared to Netherlandic Dutch, where the inflected counterpart is predominant.

In the rest of this paper, we will focus on the semantic parameter. Almost all scholars analysing the inflectional alternation cite this parameter and tend to consider it the most important explanatory factor to interpret the use of the uninflected adjective. As a matter of fact, several analyses even try to account for the use of the uninflected adjective invoking the sole semantic parameter (Honselaar, 1980; Odijk, 1992; Booij, 2002).

The semantic parameter points out that the use of the uninflected adjective identifies the NP as having non-fully compositional semantics. In such cases, the adjective-noun sequence constitutes a lexical unit that provides the name for a group of entities. The uninflected adjective *groot-∅* ('great') in *het groot-∅ seminarie* (lit. 'the great seminar') does not specify the referent of *seminarie* ('seminar') with respect to its size – as would the inflected adjective *groot-e* in *het groot-e seminarie* ('the great seminar, viz. the building'). Instead, the NP as a whole refers to an institution where boys are prepared for catholic priesthood. Booij (2002) uses the term "collocational idiom", emphasising that some meaning aspects are not predictable from the adjective or the noun, but originate from the NP construction. This is a fully fledged and productive naming mechanism in Dutch, competing with adjective-noun compounds (De Caluwe, 1990; Booij, 2002). Such adjective-noun sequences constitute conventional expressions to designate the referent in examples such as *het stoffelijk-∅ overschot* ('the mortal remains') or *het meewerkend-∅ voorwerp* ('the indirect object'). It is worth mentioning that the Dutch reference dictionary *Van Dale* (Geeraerts, 2003) lists *groot seminarie* as an adjec-

tive-noun compound, *grootseminarie*, illustrating the fuzzy boundaries between both naming mechanisms.

The semantic parameter also covers cases where the uninflected adjective realises a semantic specialisation. Whereas the NP with the inflected adjective *het Nederlands-e elftal* ('the Dutch football team') may refer to any football team with (a lot of) Dutch players, the uninflected counterpart *het Nederlands-∅ elftal* narrows the reference to the Dutch national team (Honselaar, 1980).

Summarising, according to the semantic parameter, the absence of the adjectival inflectional ending is considered to be a formal indication of semantic specialisation. The morphologically marked form, i.e. the uninflected adjective, iconically identifies a marked meaning.

A closer examination of the semantically oriented analyses of the inflectional alternation reveals three methodological peculiarities. First, the analyses invoking the non-compositional semantics of the NP as a factor motivating the use of the uninflected adjective, do not agree upon the obligatoriness of the uninflected adjectival form to identify a NP with non-fully compositional semantics. Whereas Odijk (1992: 202) states that "when the adjective is combined with the noun to form a N' [a NP with lexical semantics], it gets no *-e*", Booij (2002: 316) considers the use of the uninflected adjective rather as a tendency. Second, the examples quoted in the literature are usually highly idiomatic and tend to neglect adjective-noun combinations with transparent semantics. Finally, although the semantic parameter is considered a substantial argument to account for the use of the uninflected adjective, it has not yet been subject of an empirical analysis. Previous empirical studies of the adjectival inflection (Lebrun and Schurman-Swillen, 1966; de Rooij, 1980) did not measure the impact of this parameter, invoking seemingly insurmountable methodological difficulties to make it operational. All "fixed combinations" (de Rooij, 1980: 12) are excluded, without making explicit the criteria to decide when an adjective-noun combination is considered to be "fixed".

These observations motivate a thorough empirical analysis. This is the only way to evaluate the actual impact on the inflectional alternation of the frequently quoted semantic parameter. Furthermore, an empirical analysis would enable us to check whether the uninflected adjective is indeed obligatory in idiomatic expressions. However, the interpretation of the semantic parameter in terms of semantic specialisation or markedness, as is the case in most introspective analyses, is difficult to apply in an empirical analysis, let alone quantify. As a consequence, we will focus on another factor characteristic of idiomaticity, namely the collocational value of the adjective-noun combination. Geeraerts (1986: 134) defines lexical collocation as "idiosyncratic restrictions on the combinatorics of lexical elements". This definition ranges from expressions with an opaque meaning, such as *het hard-∅ gelag* ('the bad break'), to formally fixed expressions with a transparent meaning, such as *het blond(-e) bier* (lit. 'the blond beer'; 'the lager beer'). This approach agrees with lexical analyses, where idiomaticity is not considered as a discrete but rather as a gradient property (Cruse, 1986; Geeraerts, 1986; Van Sterkenburg, 1993).

We do not *a priori* exclude an analysis of the semantic effects mentioned above, but in this contribution we want to evaluate the explanatory power of lexical collocations. In the next section, we will present a quantitative alternative to compute the lexical attraction between the adjective and the noun in a NP by means of the log-likelihood ratio test.

3. Quantifying lexical collocations

The goal of the quantitative test is to compute the lexical attraction strength between the adjective and the noun comparing their expected and observed frequencies. The underlying assumption is that related words will tend to co-occur more often than they would by pure chance. Applied to the present topic, adjectives and nouns constituting a lexical collocation will tend to co-occur more frequently than what would be predicted using their unigram frequencies. Moreover, this approach enables the detection of usage effects, i.e. tendencies in language use structuring the grammar bottom-up (Bybee and Hopper, 2001). In brief, this quantitative approach makes a semantic effect operational by means of a combinatorial criterium based on usage effects.

Before presenting the measure to quantify the lexical attraction strength, it should be noted that lexical (Cruse, 1986) and morphological handbooks (Matthews, 1991) mention qualitative tests to identify lexical collocations. The rationale underlying the qualitative tests is to check whether the adjective-noun bigram possesses phonetic, morphological, syntactic and semantic properties commonly associated with NPs. Examples of such criteria are the absence of the common projection possibilities of the adjective and the noun, and the impossibility to substitute the adjective and the noun by (near-)synonyms. Lexical collocations share more properties with adjective-noun compounds than with common NPs. Although the application of those criteria allows the construction of a scale ranging from a high to a low collocational score, we choose not to apply it because these criteria rest on introspective and subjective judgements. Besides, it is not clear how to establish a hierarchy between the different criteria in order to resolve problems resulting from conflicting criteria (Matthews, 1991). In future research, the impact of these criteria must be accounted for.

Although we realise that it is difficult – if not impossible – to select the procedure that provides the best assessment of collocations[2], the log-likelihood ratio test (henceforth LL; Dunning, 1993) seems to be the best fit for the present purpose compared to other common collocation measures, such as Pearson's χ^2, the mutual information index (Church and Hanks, 1990) and the t-test. First, best results in identifying adjective-noun collocations in a pre-established list are achieved using LL. Second, LL performs well with sparse data (Dunning, 1993; Manning and Schütze, 2002). The other tests mentioned all appear to have some major drawback. The t-test assumes a normal distribution, which cannot be taken for granted in natural language (Manning and Schütze, 2002). As for the χ^2-test, its performance decreases with low frequency data. This objection also applies to the mutual information index, which systematically overestimates low frequencies.

The LL ratio compares the probability that the adjective (ADJ) and the noun (N) are independent (zero hypothesis) to the probability that ADJ and N are dependent (research hypothesis), by verifying whether the presence of N given ADJ is significantly different from N not given ADJ. Applied to *het groot-∅ seminarie*, this strategy amounts to verifying whether *seminarie* is significantly more frequent than any other noun after *groot-∅*, than after any other adjective. When this is the case, the sequence gets a high LL score; in the opposite case, it receives a low score.

[2] We do not use the association measure to identify or to extract potential lexical collocations from a text, but to compute the collocational strength between an adjective and a noun in a pre-established list. Consequently, we do not have to bother about the optimal span for the collocates, since the adjective and the noun are adjacent within the domain of agreement, viz. the NP (see section 2).

	ADJ	**¬ADJ**	**Σ**
N	a	b	a+b
¬N	c	d	c+d
Σ	a+c	b+d	n

Table 1. Contingency table to compute -2LL

Based on the 2x2 contingency table presented in table 1, the LL ratio is defined by means of the following formula: -2LL = 2(a ln(a) + b ln(b) + c ln(c) + d ln(d) + n ln(n) – (a+b) ln(a+b) – (c+d) ln(c+d) – (a+c) ln(a+c) – (b+d) ln(b+d)).

4. Empirical analysis

Before discussing and evaluating the results of the empirical analysis (section 4.2.), we will briefly present the corpus and the way we scrutinised it (section 4.1.).

4.1. *Gathering the data*

The data used in the empirical case-study are extracted from the *Corpus Gesproken Nederlands* ('Corpus Spoken Dutch'; henceforth CGN) (Oostdijk, 2001). This is a 5m words database of spoken Dutch[3]. CGN covers both lectal distinctions that are considered relevant for the alternation between inflected and uninflected adjectives: geographically, Netherlandic and Belgian Dutch data are gathered ; from a stylistic point of view, the corpus comprises different varieties of spoken Dutch, ranging from controlled monologues to spontaneous and informal face-tot-face dialogues and multilogues.

To compute the -2LL of an adjective-noun bigram, the following data have to be gathered: the frequency of the adjective-noun bigram under consideration, the number of bigrams containing ADJ but not N, the number of bigrams containing N but not ADJ, and the number of bigrams containing neither ADJ nor N. To this end, a list is compiled of all bigrams occurring at least three times in CGN (n = 3237915) and their respective token frequencies. All bigrams occurring less than three times in the corpus are excluded, because they would yield statistically insignificant results. Next, this overall bigram list is compared with the database of adjective-noun bigrams appearing in a NP with a singular neuter head noun and a definite determiner[4] (n = 4970), viz. the agreement domain where the inflectional variation under investigation shows up (see section 2). This matching operation yields a list of 1489 adjective-noun bigrams for which a -2LL score is computed. This -2LL value is used to rank the bigrams. The remainder of the database consists of bigrams occurring less than three times in CGN. Those bigrams (n = 3481) are assigned a -2LL value of 0, the lower limit of the -2LL ranking.

4.2. *Evaluating the results*

The output of the LL test will be processed in two stages. First, the -2LL ratio will be plotted to the distribution of inflected and uninflected adjectives in order to verify the empirical

[3] The final release (release 7) of CGN, which is expected in January 2004, will comprise 10m words. The present analysis uses release 5 (April 5, 2002), containing about 5m words.

[4] Adjectives that are always inflected, such as the superlative in attributive position (*het mooist-e paard*/**het mooist-∅ paard* 'the most beautiful horse'), or always uninflected, such as adjectives ending in a vowel (**het lila-e jasje/het lila-∅ jasje* 'the lilac jacket'), are removed from the database, since they do not exhibit the variation under investigation.

validity of lexical collocations as a parameter to interpret the alternation between inflected and uninflected adjectives. Second, the -2LL score will be entered, together with the other potential explanatory parameters mentioned in section 2, as regressor variables in a binomial logistic regression with the probability of the uninflected adjective, or more precisely the odds of the probability of the uninflected adjective, as response variable. This multifactorial statistic will be used to quantify the impact of the semantic parameter and to compare its relative strength to that of the other regressors in the model. At this end, the relevant adjective-noun bigrams (n = 4970) are analysed with respect to the other relevant parameters influencing the alternation between inflected and uninflected adjectives (see section 2)[5].

As mentioned in the previous section, only 1489 out of the 4970 adjective-noun bigrams occur at least three times in CGN. This entails that 3481 observations are assigned a -2LL score of 0. Table 2, where 'abs' stands for absolute frequency and 'rel' for relative frequency, shows that the subset of the observations with a -2LL higher than 0 contains significantly more uninflected adjectives than the subset with a -2LL equal to 0 (df = 1; $\chi^2 = 257.7166$; $p \leq 0.001$)[6].

adjective	-2LL = 0		-2LL > 0		Σ	
	abs	**rel**	**abs**	**rel**	**abs**	**rel**
inflected	2891	0.8305	924	0.6206	3815	0.7676
uninflected	590	0.1695	565	0.3794	1155	0.2324
Σ	3481	1.0000	1489	1.0000	4970	1.0000

Table 2. Distribution of both adjectival forms in bigrams with -2LL = 0 and -2LL > 0

These figures indicate a positive effect of highly recurrent bigrams on the use of the uninflected adjective. Focussing on the subset with -2LL > 0, this tendency can be analysed in further detail. First, this subset displays a positive correlation between the -2LL score and the use of the uninflected adjective as demonstrated by the positive correlation coefficients between the cumulative probability of the uninflected adjective at the one hand and the -2LL score and the -2LL ranking at the other:

- $r(Pr_{cum}(ADJ\_uninflected),LL\_score) = 0.48294$
- $r(Pr_{cum}(ADJ\_uninflected),LL\_ranking) = 0.98709$

Second, the 1489 bigrams tokens for which -2LL > 0 represent 345 bigram types. For 289 of those bigram types (83.77%), all occurrences allow one adjectival form. In other words, bigrams where the adjective and the noun co-occur more often than would be expected by pure chance manifest a strong tendency to favour the use of one inflectional alternative, either the inflected or the uninflected, to the detriment of the other. This result puts the generally accepted view that idiomaticity motivates the use of uninflected adjectives in a broader perspective. It will have to be verified whether the use of the inflected adjective in such cases is motivated by semantic or other parameters.

[5] The empirical analysis is performed by means of *Abundantia Verborum*. This is a computer tool to query corpora and to label and classify the extracted data. For more information, we refer to Speelman (1997) or http://wwwling.arts.kuleuven.ac.be/genling/abundant.

[6] The subset with -2LL = 0 shows a very homogeneous distribution of both adjectival alternatives as indicated by the low variance of the cumulative probabilities of the uninflected adjective ($\sigma^2 = 0.000426$).

A closer look at the data reveals that no inflected adjective appears amongst the 196 tokens or 11 types with the highest -2LL values. Of the 50 bigram types (417 tokens) at the top of the -2LL ranking, only three bigram types have a highly idiomatic meaning, but remarkably, the adjective takes an *-e* suffix in all three cases: *het Oud-e Testament* ('The Old Testament'), *het Groen-e boekje* (lit. 'the little green book', hence 'the standard word list of Dutch) and *het groen-e Poldermodel* (lit. 'the green Polder model', hence 'the political model for polder ecology'). The other bigrams in this top list are lexical collocations, such as *het Europees-∅ kampioenschap* ('the championship for European national teams') or *het cultureel-∅ centrum* ('the cultural centre'). Amongst these sequences, we find conventional names for official institutions, such as *het openbaar-∅ vervoer* ('the public transport') or *het secundair-∅ onderwijs* ('secondary education').

Summarising, compared to adjective-noun combinations with a loose association strength, sequences with a strong association between adjective and noun seem to favour the use of the uninflected adjective. Moreover, such bigrams show little internal variation between both inflectional alternatives, revealing the importance of usage effects. On the other hand, the data clearly show that the use of the uninflected adjective is not a binary but a gradient matter. In order to gain insight in the effective impact of the semantic parameter, we will enter it as a regressor variable in a logistic regression together with the factors mentioned in section 2. At this point, it is important to mention that the resulting regression model is not the end of the research. The other parameters need further refinement and extension.

A logistic regression analysis[7] (Rietveld and van Hout, 1993: chap. 9; Paolillo, 2002: chap. 8) allows the use of categorical and numeric variables. Moreover, since an additive model may predict values outside the range [0,1], the dependent variable is transformed in a logit function, compelling the use of a non-linear statistic, viz. the logistic regression. This statistic computes the best fitting logistic curve for a given distribution applying maximum likelihood estimation to determine the coefficients of the regressor variables. We will use the logistic regression not only to quantify the impact of lexical collocations on the inflectional alternation, but also to estimate its relative impact compared to other potential explanatory factors quoted in the literature by comparing their respective estimates. Note that the logistic regression calculates changes in the log odds of the dependent variable, not changes in the dependent variable itself. Therefore, standardised estimates are calculated, the so-called β weights. The β weights of the factor values are interpreted as follows: a β weight equal to 5 for a factor value indicates that the odds ratio uninflected vs. inflected is predicted to increase 5 times as a result of the effect of that factor. Conversely, a β weight equal to 0.2 means that the odds ratio uninflected vs. inflected is predicted to decrease 5 times as a result of the effect of that factor value.

The potential explanatory factors mentioned in section 2 are made operational as follows:

- part of speech of the determiner: definite article (*het* 'the'), demonstrative pronoun, possessive pronoun, genitive determiner
- adjectival length: 1 syllable, 2 syllables, 3 syllables, more than 3 syllables
- accentuation of final adjectival syllable: not accentuated, accentuated
- register: monologue, dialogue/multilogue
- region: Netherlandic Dutch, Belgian Dutch
- lexical collocation: -2LL of the adjective-noun bigram

[7] The regression analysis has been conducted using the R statistical package. More information can be found on http://www.r-project.org/.

Note that only the -2LL factor is numeric. For the categorical factors, the first value is the control value to which the other values are compared in order to determine their effect on the response variable. The variables are added to the model by means of the forward stepwise selection method. This selection procedure starts with an empty model and adds at each stage the best performing regressor variable to the model, namely the regressor variable performing the strongest reduction of the unexplained variation. Table 3 outlines the building of the regression model applying the forward stepwise selection procedure. The unexplained variation is computed by means of AIC[8].

Step	AIC	Regression model
Start	5391	
Step 1	4697.02	response.variable = -2LL.ADJ.N
Step 2	4093.97	response.variable = -2LL.ADJ.N + adj.length
Step 3	3974.37	response.variable = -2LL.ADJ.N + adj.length + register
Step 4	3881.96	response.variable = -2LL.ADJ.N + adj.length + register + adj.final.accent
Step 5	3821.30	response.variable = -2LL.ADJ.N + adj.length + register + adj.final.accent + region
Step 6	3815.19	response.variable = -2LL.ADJ.N + adj.length + register + adj.final.accent + region + determiner

Table 3. Building the regression model using the stepwise forward procedure

Since the -2LL ratio is the first parameter to be entered to the regression model, it can be considered the best explaining factor for the use of the uninflected adjective. The data in table 3, however, need further refinement. Therefore, let's turn to table 4. This table summarises the impact of the different factor values included in the logistic regression analysis.

regressor	estimate	β weight	95% CI β weight	
			Lower	Upper
(intercept)	-3.2139			
det.demonstrative	-0.2392	0.7873	0.5975	1.0373
det.possessive	0.3707	1.4488	1.1223	1.8703
det.genitive	0.1820	1.1996	0.3419	4.2096
adj.length.2syl	0.8056	2.2381	1.6778	2.9855
adj.length.3syl	1.9844	7.2745	5.6037	9.4435
adj.length.>3syl	2.0927	8.1070	5.9808	10.989
adj.final.accent.yes	-0.9282	0.3953	0.3290	0.4749
register.dial/multil	1.0433	2.8386	2.4008	3.3562
region.belgium	0.6718	1.9577	1.6541	2.3170
-2LL.ADJ.N	0.0101	1.0102	1.0089	1.0114

Table 4. Result of the logistic regression analysis

First, we will discuss the effect of the individual factor values in the model. Looking at the confidence intervals, all regressors, except the demonstrative pronoun and the genitive determiner, appear to differ in a significant way from the control value. The effect of these regres-

[8] AIC (Akaike Information Criterion) is an adjusted measure to compute the (un)explained variation in a fitted model. Due to a bias, being a function of the number of degrees of freedom in the model, maximum likelihood estimation is not suited to compare different fitted models. This bias is adjusted by AIC.

sors does not significantly differ from the definite article *het* ('the'). Moreover, the factor values corroborate the tendencies described in section 2. Comparing the impact of the different factors in the regression model, the -2LL ratio is the best performing parameter. It is not only the first parameter to be added to the stepwise regression model as shown in table 3, but it also has the greatest effect on the predictive power of the model. Although its coefficient estimate and β weight are relatively small compared to those of other significant values, it should be kept in mind that this factor is not binary (0 vs. 1), as is the case for the categorical factor values, but numeric, with a mean value of 60.65.

Next, we will assess the quality of the fitted model examining its predictive and explanatory power. When used to predict the adjectival inflection in the 4970 observations in the database, the fitted model achieves a success rate of 83.76%. This score is an important improvement with respect to the intercept only model, which has a success rate of 76.76%. Let's turn to the explanatory power of the fitted model. This can be assessed looking at the variation explained by the fitted model, which is the remainder of the total variation in the intercept only model minus the unexplained variation in the fitted model: 5389.00 – 3793.18 = 1595.82. This figure yields a model χ^2 $p \leq 0.001$ (df model $\chi^2 = 10$). Hence, we may conclude that the fitted model reduces the unexplained variation in a highly significant way.

The logistic regression analysis shows that the semantic parameter is the most important factor to account for the use of the uninflected adjective. Nevertheless, it is not a sufficient factor, as indicated by the significant impact of other parameters and the success rate of the regression analysis which is inferior to 100%. We also have to bear in mind that we used a combinatorial criterium, namely the attraction strength within lexical collocations, to test the semantic effect, in contrast to introspective analyses, which mostly rely on idiomaticity. Furthermore, the effects of the other parameters included in the regression model need to be analysed in further detail.

5. Conclusions

In this paper, we have presented an approach to empirically validate semantic effects in a case study of variational linguistics, namely the alternation between inflected and uninflected Dutch attributive adjectives. The semantic effect of idiomaticity was made operational in terms of lexical collocations, looking for co-occurrence patterns within adjective-noun bigrams. The notion of lexical collocation was argued to seize the insight that idiomaticity is not a binary, but a gradient notion. Furthermore, lexical collocations can be quantified in a straightforward way, computing the lexical attraction between the adjective and the noun.

The empirical analysis revealed a clear influence of lexical collocations on the inflectional variation, favouring the use of the uninflected variant when the adjective and the noun co-occur more frequently than would be predicted by pure chance. Furthermore, tightly associated adjective-noun combinations showed little internal variation between both adjectival alternatives. In order to seize the real influence of lexical collocations on the inflectional alternation, we entered it, together with other relevant factors, in a logistic regression analysis. This analysis corroborated the widespread opinion amongst Dutch linguists that the semantics of the adjective-noun combination is the predominant motivation to use the uninflected adjective.

Different issues remain to be tackled in future research. The quantitative results of the collocation analysis must be completed with a qualitative interpretation, focussing on other, more semantic characteristics of idiomaticity. With respect to the bigram types allowing only one

inflectional variant, elements determining the choice for the given alternative must be looked for. As for the regression analysis, the regressor variables have to be refined and extended in order to increase the explanatory power of the model.

References

ANS (1997). Haeseryn W., Romijn K., Geerts G., de Rooij J. and van den Toorn M.C. *Algemene Nederlandse Spraakkunst*. Martinus Nijhoff – Wolters Plantyn.

Cruse D.A. (1986). *Lexical Semantics*. Cambridge University Press.

Booij G. (1992). Congruentie in Nederlandse NP's. *Spektator*, vol. (21/2): 119-135.

Booij G. (2002). Constructional Idioms, Morphology, and the Dutch Lexicon. *Journal of Germanic Linguistics*, vol. (14/4): 301-329.

Bybee J. and Hopper P. (Eds) (2001). *Frequency and the emergence of linguistic structure*. John Benjamins.

Church K.W. and Hanks P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, vol. (16): 22-29.

De Caluwe J. (1990). Complementariteit tussen morfologische en in oorsprong syntactische benoemingsprocédés. In De Caluwe J. (Ed.), *Betekenis en Productiviteit* (*Studia Germanica Gandensia*). *Seminarie voor Duitse taalkunde*: 9-24.

Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, vol. (19/1): 61-74.

Evert S. and Krenn B. (2001). Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, July 6-11: 188-195.

Geeraerts D. (1986). *Woordbetekenis: een overzicht van de lexicale semantiek*. Acco.

Geeraerts D. (Ed.) (2003). *Van Dale Groot woordenboek der Nederlandse taal*. CD-ROM edition, version 1.3. Van Dale Lexicografie.

Honselaar W. (1980). On the semantics of adjective-noun combinations. In Barentsen A.A., Groen B.M. and Sprenger R. (Eds), *Studies in Slavic and General Linguistics*, vol. (1): 187-206.

Lebrun Y. and Schurmans-Swillen G. (1966). Verbogen tegenover onverbogen adjectieven in de taal van de Zuidnederlandse dagbladpers. *Taal en Tongval*, vol. (18/1): 175-187.

Manning Chr. and Schütze H. (2002). *Foundations of statistical natural language processing*. MIT Press.

Matthews P.H. (1991). *Morphology*. Cambridge University Press.

Odijk J. (1992). Uninflected Adjectives in Dutch. In Bok-Bennema R. and van Hout R. (Eds), *Linguistics in the Netherlands*. John Benjamins: 197-208.

Oostdijk N. (2000). Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, vol. (5/3): 280-284.

Paolillo J.C. (2002). *Analyzing Linguistic Variation. Statistical Models and Methods*. CSLI Publications.

Rietveld T. and van Hout R. (1993). *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter.

Rooij J. de (1980). Ons bruin(e) paard. *Taal en Tongval*, vol. (32): 3-25: 109-129.

Speelman D. (1997). *Abundantia Verborum. A computer tool for carrying out corpus based linguistic case studies.* PhD dissertation, KU Leuven.

Van Sterkenburg P. (1993). Gelexicaliseerde woordgroepen van het type A+N. *Tabu*, vol. (23/1-2): 131-142.

La “malattia mentale” tra senso scientifico e senso comune: analisi di testi di settore e divulgativi

Gian Piero Turchi, Sara Martinazioli, Luisa Orrù, Barbara Laliscia

Dipartimento di Psicologia Generale – Università di Padova
Via Venezia, 8 – 35131 Padova – Italy
gianpiero.turchi@unipd.it, orru@psy.unipd.it

Abstract

The aim of the present work is the investigation of the discursive productions whose object is the “mental illness”, through textual analysis of mass communication texts and sectorial publications of psychiatry and clinical psychology. The basic theoretical assumption is that the objects of knowledge are not independent from the discourses that are generated around them (Gergen, 1985). The textual material choosen for this research is composed of 226 articles divided in two groups: 176 articles were selected from divulgative publications, while 50 articles were choosen among sectorial publications. The analysis of lexical correspondences has been then applied to the two corporas, using the software SPAD-t (Bolasco, 1999).
In regard to “divulgative publications”, a diffused presence of matters of “psychiatric” and “psychological” nature has been observed, not only in specific publications dedicated to the divulgation of scientific issues, but also in the daily press and in a lot of journalistic headings. In regard to “sectorial publications”, some fundamental themes can be pointed out: the main thematic observed can be ascribed to an “empiricist” and “positive” model of knowledge; furthermore, two separate thematic groups have emerged, one belonging to the research level, the other to the operational level and therefore to the clinical level. It was observed, instead, how much the scientific knowledge normally affects the constructions of common sense, contributing to modify or consolidate them, as well as how much the social representations (the common sense) affect the construction of the scientific sense.

Riassunto

Il presente lavoro ha come obiettivo l’indagine delle produzioni discorsive che hanno per oggetto la “malattia mentale”, attraverso l’analisi testuale di testi di comunicazione di massa e delle pubblicazioni di settore di psichiatria e di psicologia clinica. L’assunto teorico di base è che gli oggetti di conoscenza non siano indipendenti dai discorsi che vengono prodotti intorno ad essi (Gergen, 1985). Il materiale testuale scelto per questa ricerca è composto da 226 articoli suddivisi in due gruppi: 176 articoli sono stati selezionati da testate divulgative, mentre 50 articoli sono stati individuati tra le pubblicazioni di settore. Ai due corpora è stata applicata l’analisi delle corrispondenze lessicali mediante software SPAD-t (Bolasco, 1999).
Per quanto riguarda i “testi divulgativi” si è potuto constatare la presenza diffusa di argomenti di carattere “psichiatrico” e “psicologico” rintracciabile, non solo in pubblicazioni specifiche per la divulgazione di argomenti scientifici, ma anche nella stampa quotidiana e in molte testate giornalistiche. Per quanto riguarda i “testi di settore” si possono individuare alcuni punti fondamentali: la principale tematica individuata si rifà essenzialmente ad un modello di conoscenza di tipo “empirista” e “positivo”; inoltre sono emersi due gruppi tematici distinti, l’uno appartenente all’ambito della ricerca, l’altro all’ambito dell’intervento e dunque della clinica. In generale si può evidenziare invece come la conoscenza scientifica incida sulle costruzioni di senso comune, contribuendo a modificarle o a consolidarle, così come le rappresentazioni sociali (il senso comune) influiscano sulla costruzione del senso scientifico.

Keywords: textual data analysis, divulgative publications, sectorial publications, mental illness, discursive processes.

1. Introduzione

Il presente lavoro si propone di indagare le produzioni discorsive che hanno per oggetto la "malattia mentale", utilizzando l'analisi del contenuto di testi prodotti nell'ambito delle comunicazioni dei *mass media* e delle pubblicazioni di settore relative alla psichiatria e alla psicologia clinica. L'assunto teorico di base è che gli oggetti di conoscenza non siano indipendenti dai discorsi che vengono prodotti intorno ad essi (Gergen, 1985), anzi sono le procedure conoscitive, equiparate a produzioni discorsive, che costruiscono e configurano gli stessi oggetti d'indagine (Turchi, 2000). Pertanto *focus* della presente ricerca divengono sia i processi di conoscenza (*ergo* i processi discorsivi), considerati come attività sociale condivisa, attraverso cui si costruisce e si struttura la realtà (da qui la scelta relativa alla tipologia di testi), sia gli strumenti di diffusione della conoscenza che riproducono e consolidano particolari configurazioni della realtà stessa (da qui la scelta relativa alla tipologia di fonti da cui trarre i testi da sottoporre ad analisi). Sulla base di questa prospettiva la "malattia mentale" viene considerata un costrutto teorico, quindi una "formazione culturale", storicamente e socialmente situata ed utilizzata all'interno delle stesse pratiche sociali che contribuiscono a costituirlo (Berger e Luckmann, 1966 ; Salvini, 1988 e 1998). In particolare il costrutto "malattia mentale" risulta un processo discorsivo che storicamente risulta assorbito ad un orizzonte medico-scientifico (Foucault, 1961), caratterizzandosi così come una categoria astratta, descritta da polarizzazioni come normalità e patologia, salute e malattia (Canguilheim, 1966). Questo ha fatto si che, relativamente a tale costrutto le possibilità per l'uomo comune di esplorare il mondo scientifico, passino solo attraverso canali di divulgazione di massa come la stampa o i programmi televisivi. Il risultato di questo processo storico è stata l'individuazione: da una parte di una comunità ristretta di specialisti del sapere che si rende sempre più autonoma; dall'altra di una cultura di massa non più basata sul sapere popolare e tradizionale, bensì su nozioni semplificate derivanti dal sapere scientifico (Moscovici, 1984). In questo modo sembra pertanto che, nel caso della "malattia mentale", sia la conoscenza scientifica ad incidere sulle costruzioni di senso comune, contribuendo a modificarle o a consolidarle; di converso, quanto le costruzioni di senso comune influiscano sul senso scientifico, appare assai meno scontato e soprattutto non oggetto d'indagine. In questa sede invece si intende prendere in considerazione anche questo aspetto.

Per cui proprio in virtù di quanto delineato, questo lavoro intende evidenziare quali siano i reciproci rapporti di influenza tra il senso comune e il senso scientifico, che in termini di processi discorsivi, concorrono alla costruzione dell'oggetto di conoscenza "malattia mentale".

2. Metodologia

Nonostante il progresso tecnologico abbia visto come protagonista la metodologia dell'analisi del testo, con l'applicazione di sempre più sofisticati strumenti di ricerca (SPAD-T, LEXICO3, HYPERBASE, Alceste, Atlas.ti, Taltac, ecc.), rimangono attuali gli oggetti d'indagine che coinvolsero Lasswell (Losito, 1993) fin dagli anni '50. Il presente lavoro si pone come obiettivo quello di evidenziare le pratiche discorsive attraverso cui i testi di divulgazione scientifica e di comunicazione di massa strutturano l'oggetto di conoscenza "malattia mentale". Si sono pertanto sottoposti ad analisi testuale due gruppi di testi: pubblicazioni di carattere divulgativo ad argomento scientifico, testate giornalistiche e quotidiani in generale; pubblicazione di settore in ambito psichiatrico e in ambito clinico. I criteri di selezione dei testi oggetto d'indagine, hanno riguardato il periodo di pubblicazione e l'argomento dell' articolo. Gli articoli scelti sono databili in un periodo di tempo che va dal 1999 al 2002, ponendo quindi l'attenzione su pubblicazioni recenti. Per quanto riguarda l'argomento sono

stati selezionati sia articoli che si occupano di "disturbi mentali" in generale, sia articoli che si concentrano su specifici "disturbi psichici", riferibili alle classificazioni diagnostiche del DSM-IV. La maggior parte degli articoli dei testi divulgativi è stata recuperata tramite internet grazie ai motori di ricerca propri delle riviste che fornivano gli articoli integralmente.

2.1. Procedure d'analisi

Il materiale che di seguito verrà presentato descrive i risultati ottenuti dall'applicazione dell'analisi delle corrispondenze lessicali mediante software SPAD-t ed è stata applicata ai due corpora presi in esame. È possibile pertanto costruire una tabella di contingenza con in riga le forme grafiche (una forma grafica corrisponde ad una riga) e in colonna i testi (un testo, in questo caso un articolo, corrisponde ad una colonna) (Bolasco, 1999). Le forme grafiche sono considerate come modalità della variabile "Lessico", incrociata con la variabile "Testo", che presenta tante modalità quanti sono i testi analizzati (in questo caso per il corpus "testi divulgativi" le modalità sono 176, mentre per il corpus "testi di settore" 50 modalità). In questo modo si può descrivere come le parole si associno ai diversi testi e come i differenti testi si associno alle parole in essi utilizzate. Pertanto è stata prescelta la procedura Asparm per effettuare le analisi delle corrispondenze sui due corpora testuali oggetto della ricerca.

2.2. Descrizione dei corpora

Il materiale testuale scelto per questa ricerca è composto da 226 articoli suddivisi in due gruppi: 176 articoli sono stati selezionati da testate di divulgazione di massa, mentre 50 articoli sono stati individuati tra le pubblicazioni di settore di psichiatria e psicologia clinica. Pertanto l'oggetto dell'analisi testuale, che di seguito verrà descritta, è costituito da due corpora: "testi divulgativi" e "testi di settore". Sulla base degli obiettivi della ricerca, si è preferito non creare un unico corpus testuale composto da tutti gli articoli selezionati, in modo da costituire profili lessicali distinti che andranno poi a definire la base per le successive procedure di analisi. Questo tipo di scelta vuole tenere in considerazione le differenze linguistiche, strutturali e funzionali di questi due gruppi di testi ed in più ha permesso di ottenere degli indici di confronto tra queste due produzioni linguistiche.

2.2.1. *I vocabolari*

Dal corpus "testi divulgativi" (N = 100.382) si è ottenuto un vocabolario di 16.754 (V) forme grafiche distinte. Dal corpus "testi di settore" (N = 161.800) si è ottenuto un vocabolario di 18.903 (V) forme grafiche distinte. Per quanto riguarda i corpora dei testi della presente ricerca gli indici di ricchezza lessicale (Tuzzi, 2003, p. 76) sono rispettivamente di 16% per il corpus "testi divulgativi" e di 11% per il corpus "testi di settore". Il corpus più esteso ("testi di settore") presenta un indice di ricchezza lessicale inferiore al corpus meno esteso ("testi divulgativi"), indicando quindi una maggiore omogeneità lessicale dei testi di settore rispetto a quelli divulgativi.

2.2.2. *Riduzione di soglia*

Per entrambi i corpora, oggetto di indagine della presente ricerca, si è scelta una soglia di frequenza pari a 6, in modo da considerare nelle analisi statistiche le classi di frequenza da 7 in su. È possibile mettere in luce che per il corpus "testi divulgativi" con un taglio di frequenza di 6, si è ottenuto una copertura del corpus testuale del 73%, e del vocabolario del 12%, mentre per il corpus "testi di settore" si è ottenuto una copertura del testo di 82% e del vocabolario di 15%. Si sono così effettuate due procedure Asparm, una per il corpus "testi

divulgativi", sottoponendo ad analisi complessivamente 1644 forme grafiche distinte, l'altra per il corpus "testi di settore" sottoponendo ad analisi 2349 forme grafiche distinte.

3. Esposizione e discussione dei dati

3.1. *Criteri seguiti per l'esposizione e la discussione dei risultati*

Nelle seguenti analisi vengono presi in considerazione i fattori primo e secondo, sia rispetto ai "testi divulgativi", sia riguardo ai "testi di settore", e i piani fattoriali originati dal loro incrocio, in quanto spiegano la maggior parte dell'inerzia totale e si differenziano in maniera netta dai rimanenti fattori.

numero	valore proprio	percentuale	% cumulativa	
1	0,22	1,68	1,68	**********
2	0,21	1,59	3,28	**********
3	0,21	1,58	4,86	**********

Tabella 1. Istogramma dei primi 3 fattori relativi al corpus "testi divulgativi".

numero	valore proprio	percentuale	% cumulativa	
1	0,22	4,66	4,66	**********
2	0,21	4,34	9,00	**********
3	0,19	3,96	12,96	**********

Tabella 2. Istogramma dei primi 3 fattori relativi al corpus "testi di settore".

Per migliorarne la leggibilità, nei grafici si riportano solo le forme lessicali che hanno contribuito in maniera significativa alla costruzione di almeno uno dei due fattori, e la cui inerzia risulta a sua volta significativamente spiegata da essi; inoltre sono stati proiettati sul piano fattoriale le forme grafiche in relazione alla categoria grammaticale di appartenenza (sostantivi, verbi e aggettivi-avverbi). Nella descrizione che segue verranno illustrati esclusivamente i piani fattoriali relativi ai sostantivi. La vicinanza tra due o più elementi viene considerata indice di una possibile associazione; nel caso di contiguità tra le forme lessicali, diviene possibile rilevare l'*area semantica* cui un certo numero di termini fa riferimento (Bolasco, 1999).

3.2. *Analisi del piano fattoriale definito dai fattori 1 e 2 "testi divulgativi"*

Nel piano fattoriale definito dall'incrocio tra i primi due fattori (percentuale cumulativa di inerzia spiegata 3,28) si evidenziano alcune aree semantiche di seguito descritte. Relativamente al primo fattore, sul semiasse positivo, si colloca la polarità "psichiatrica", e sul semiasse negativo, la polarità "scientifica". Il secondo fattore si suddivide nella polarità della "pratica medica", nel semiasse positivo, e nella polarità "giuridico-morale", nel semiasse negativo. Nel primo quadrante si configura un'area semantica caratterizzata da repertori riguardanti l'ospitalizzazione e la malattia mentale, caratterizzata infatti da forme grafiche di tipo medico e sanitario. Il tema che emerge è relativo alla pratica collettiva di organizzazione sanitaria dell'assistenza al "malato di mente". La malattia mentale si configura come

problema sanitario che riguarda sia la società sia le istituzioni. Si trovano per esempio, forme grafiche quali “assistenza” (C.a. 0.7 F2), “malati-mente” (C.a. 0.6 F1, F2), “ex-degenti” (C.a. 0.4 F1, F2), “posti-letto” (C.a. 2,1 F2) “ministro” (C.a. 0.3 F1), “la-legge” (C.a. 0.5 F1), “salute-mentale” (C.a. 0.2 F2), “territorio” (C.a. 0.2 F2, F1), “asl” (C.a. 0.4 F2), “oms” (C.a. 0.7 F2). Inoltre in questo quadrante, si trovano i discorsi intorno ai disturbi del comportamento alimentare, nel caso scientifico fortemente legati ad una concezione ospedaliera del problema “anoressia-bulimia” (C.a. 0.3 F2), “disturbo-comportamento-alimentare” (C.a. 0.8 F2). Proseguendo nel secondo e terzo quadrante della rappresentazione grafica del piano fattoriale si individuano due aree semantiche, una che rintraccia repertori relativi ai vari tipi di trattamento e alla loro efficacia, ed una che configura invece una tematica relativa alla conoscenza.

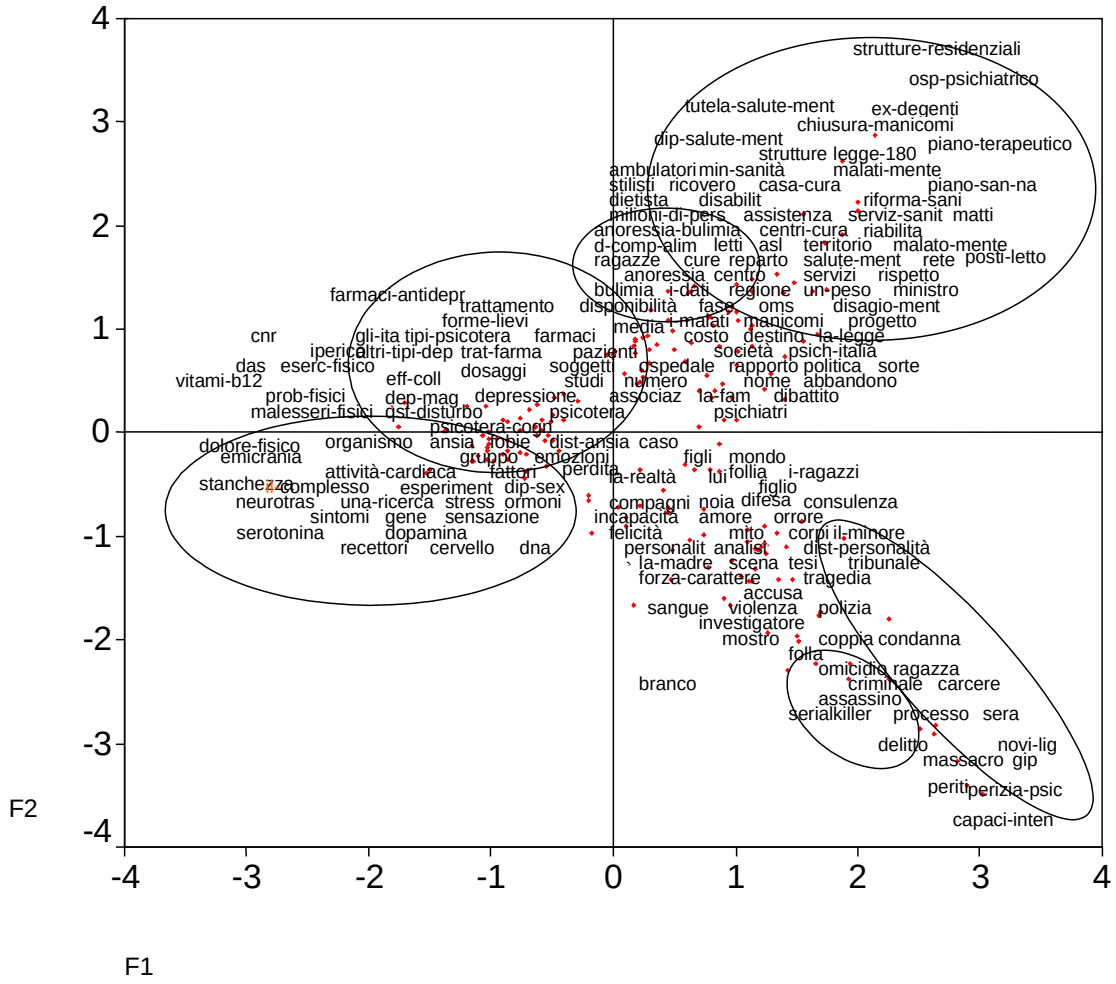


Figura 1. Rappresentazione grafica del piano fattoriale definito dai fattori uno e due. Corpus “testi divulgativi”. Soglia di contributo > 10.
Categoria grammaticale: sostantivi.

Difatti nel secondo quadrante l’area semantica individuata è caratterizzata da forme grafiche quali “psicoterapia” (C.a. F1), “farmaci” (C.a. 0.2 F1), “trattamento” (C.a. 0.2 F1), “dosaggio” (C.a. 0.2 F1), “tipi-psicoterapie” (C.a. 0.2 F1). Nel terzo quadrante invece emergono forme quali “esperimento” (C.a. 0.2 F1), “una-ricerca” (C.a. 0.2 F1), “gruppo” (C.a. 0.2 F1); forme che indicano il substrato biochimico ed organico come “gene” (C.a. 0.3 F1), “dna”

(C.a. 0.2 F2), “cervello” (C.a. 0.5 F1), “organismo” (C.a. 0.3 F1), “ormoni” (C.a. 0.2 F1), “neurotrasmettitori” (C.a. 0.3 F1); forme che indicano i vari aspetti dei disturbi, come “fattori” (C.a. 0.2 F1), “sintomi” (C.a. 0.5 F1), “sensazione” (C.a. 0.2 F1), “emozioni” (C.a. 0.2 F2) e forme grafiche che riguardano la fisicità dei disturbi, quali “dolore-fisico” (C.a. 0.2 F1), “emicrania” (C.a. 0.2 F1), “stanchezza” (C.a. 0.2 F1). Quest’area semantica mette in luce un repertorio di tipo “organico” e “fisico”, configurando quindi una visione della ricerca riconducibile al modello “materialista” e “riduzionista”; nella stessa sembra quindi configurarsi una doppia istanza della conoscenza scientifica: l’una che riguarda la necessità della spiegazione e comprensione dei fenomeni, l’altra che obbedisce ad una necessità operativa, di intervento. Il quarto quadrante raccoglie l’area semantica giuridica e morale diametralmente opposta all’area semantica relativa alla terapia. La prima raccoglie forme grafiche come “perizia-psichiatrica” (C.a. 2.3 F2), “capacità-intendere-volere” (C.a. 2 F2), “periti” (C.a. 1.8 F2), “condanna” (C.a. 0.4 F1, F2); la seconda è costituita da forme quali “criminale” (C.a. 0.8 F2), “assassino” (C.a. 0.8 F2). Da una parte emerge quindi la necessità di conoscere e giudicare e dall’altra quella di curare. L’immagine proposta è quindi quella di una scienza oggettiva e di una cura certa ed efficace.

3.3. *Analisi del piano fattoriale definito dai fattori 1 e 2 “testi di settore”*

Nel piano fattoriale ottenuto dall’incrocio tra il primo fattore con il secondo (percentuale cumulativa di inerzia spiegata 3,82) è possibile individuare le polarità che costituiscono gli assi fattoriali: per il primo fattore la polarità positiva è “misurazione del comportamento” *versus* la polarità negativa “fenomenologica”; per il secondo fattore la polarità positiva è “naturalizzazione del comportamento” *versus* la polarità negativa “clinica prognostico-evolutiva”. Lo spazio corrispondente al primo quadrante (Fig.2) è configurato da un’area semantica riguardante il dominio sperimentale; emergono forme grafiche quali: “gruppo-controllo” (C.a. 1.1 F1), “scala” (C.a. 1.6 F1), “punteggio-medio” (C.a. 0.2 F1), “differenza-statistica-significativa” (C.a. 0.6 F1), “punteggio” (C.a. 1.6 F1). Inoltre è rintracciabile un’ulteriore area semantica individuata dalle forme grafiche riconducibili a possibili indici di misurazione dei vari ambiti della vita di una persona, quali “ostilità” (C.a. 0.3 F2), “impulsività” (C.a. 0.2 F2), “aggressività” (C.a. 0.2 F2), “trauma” (C.a. 0.2 F2). In senso generale un aspetto interessante da considerare entro quest’area semantica è la posizione delle forme “gioco” (C.a. 0.4 F2) e “gioco-azzardo” (C.a. 0.3 F2). Quest’area semantica è caratterizzata da repertori riguardanti le spiegazioni del comportamento tramite l’ “emotività” e le “caratteristiche temperamentali”, favorendo una rappresentazione del giocatore d’azzardo come fortemente legato a questi aspetti. Collocate nel primo quadrante appaiono inoltre forme grafiche, riconducibili ad un’unica area semantica, volte ad indicare costrutti psicologici e comportamentali: “dipendenza-ricompensa” (C.a. 1 F1), “ricerca-novità” (C.a. 0.9 F1), “mascheramento-emozioni” (C.a. 0.5 F1). L’area individuata è relativa alla tematica della natura e della misura, che si rifà quindi ad un orizzonte di significati di tipo “empirista” e “realista” sia in senso metodologico che ontologico. In queste aree semantiche prevale una tematica deterministica dove il soggetto è condizionato dall’“emotività” e dalla “personalità”. Il quarto quadrante configura un’area semantica in cui prevalgono gli aspetti quantitativi e soprattutto temporali della malattia. Infatti le forme che in tale posizione emergono sono: “età-esordio” (C.a. 0.9 F2), “esordio” (C.a. 1.2 F2), “ricovero” (C.a. 1.3 F2), “evoluzione” (C.a. 1.2 F2), “esordio-tardivo” (C.a. 0.2 F1), “esordio-precoce” (C.a. 0.4 F1, F2).

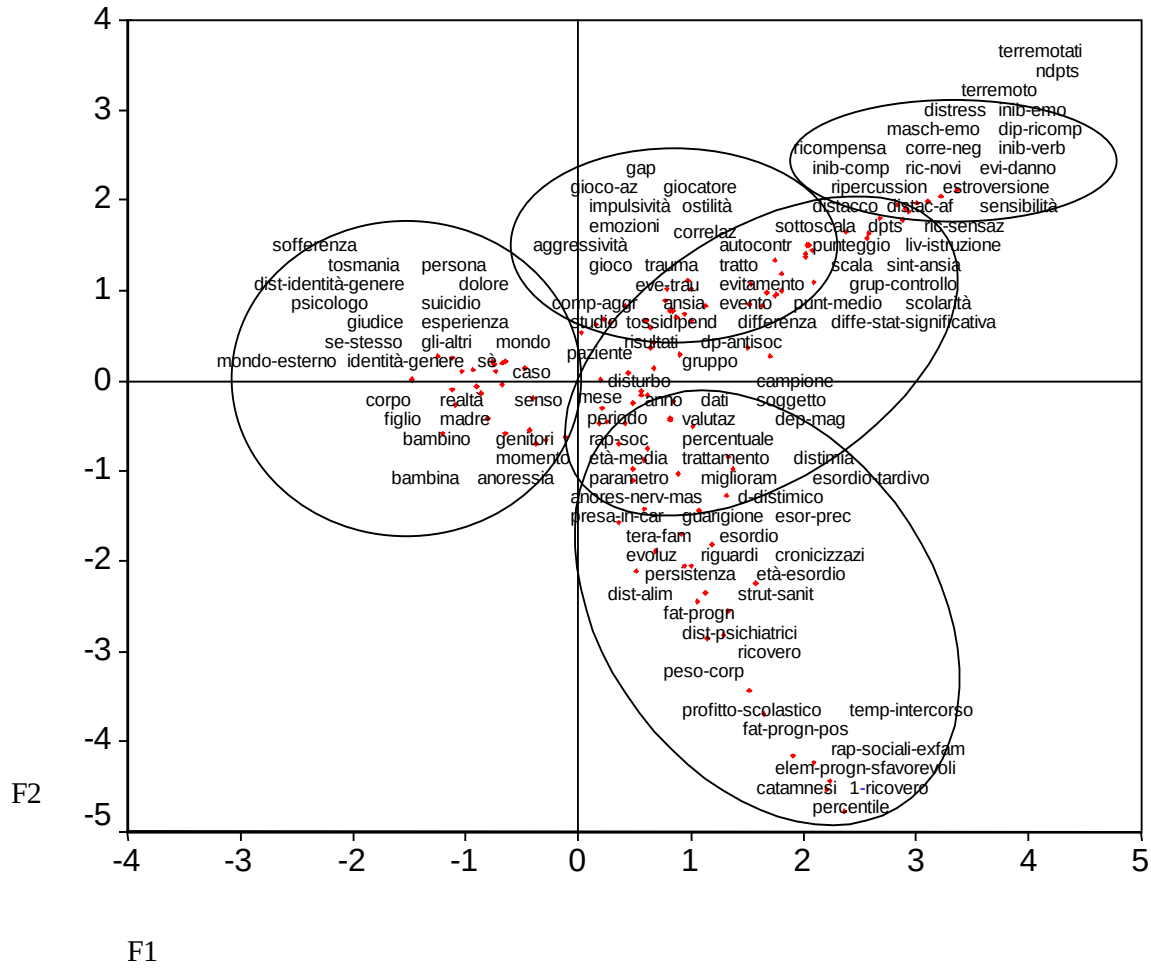


Figura 2. rappresentazione grafica del piano fattoriale definito dai fattori uno e due. Corpus "testi di settore". Soglia di contributo assoluto > 10. Categoria grammaticale: sostantivi.

Questa tematica si traduce in repertori discorsivi riguardanti l'identificazione della malattia e la possibilità di guarigione e trattamento come fortemente legati alla corretta e minuziosa descrizione della successione temporale della malattia stessa. L'area semantica che si colloca nel secondo e terzo quadrante configura un tema "fenomenologico". Sono tutte forme che indicano una condizione di sofferenza legata soprattutto all'esistenza e alla unicità della dimensione umana. Nel secondo quadrante prevale una tematica personale di esperienza e sofferenza legata al divenire storico e biografico. Si trovano infatti forme grafiche quali "persona" (C.a. 0.5 F1), "sé" (C.a. 0.3 F1), "mondo" (C.a. 0.2 F1), "gli-altri" (C.a. 0.2 F1), "dolore" (C.a. 0.2 F1), "sofferenza" (C.a. 0.6 F1), "suicidio" (C.a. 0.3 F1).

4. Conclusioni

Considerando gli obiettivi e le premesse epistemologiche che stanno alla base della presente ricerca, si è cercato di descrivere, con l'aiuto di una tecnica di analisi dei dati testuali, le tematiche e i repertori discorsivi che configurano e riproducono i discorsi relativi alla "malattia mentale". Si possono dunque proporre le seguenti riflessioni. Per quanto riguarda i "testi divulgativi" si è potuto constatare la presenza diffusa di argomenti di carattere "psichiatrico" e "psicologico" rintracciabile in molte testate giornalistiche, non solo in pubblicazioni

specifiche per la divulgazione di argomenti scientifici, ma anche nella stampa quotidiana. È rilevante sottolineare il ruolo che i mezzi di comunicazione di massa possono avere nella riproduzione di particolari repertori discorsivi, sia contribuendo così a consolidare e legittimare particolari configurazioni di realtà relative alla "malattia mentale", sia contribuendo a fare diventare le questioni di ambito psicologico e psichiatrico argomento di opinione pubblica. La visione proposta della "malattia mentale" è ancorata a repertori medici e psichiatrici sia per quanto riguarda la sua gestione pratica, sia per quanto riguarda la sua concettualizzazione teorica. Per quanto riguarda i "testi di settore" si possono individuare alcuni punti fondamentali: in primo luogo, la principale tematica conoscitiva individuata poggia essenzialmente su un modello di conoscenza di tipo "empirista" e "positivo", in secondo luogo, i repertori relativi alla pratica conoscitiva sono aderenti al modello delle scienze naturali. Sono stati infatti individuati repertori caratterizzati da processi di naturalizzazione e sostantivizzazione delle espressioni linguistiche e dei costrutti concettuali utilizzati, che sono anche rintracciabili nelle procedure di conoscenza del senso comune. Per ultimo sono state individuate due tematiche che si riferiscono a pratiche cliniche diverse che si configurano attraverso modalità discorsive differenti. Una è relativa agli aspetti di classificazione diagnostica, in base alla quale l'aspetto terapeutico è particolarmente legato alla corretta individuazione dell'etichetta diagnostica, viceversa l'altra è definita come "evolutivo-prognostica", caratterizzata da repertori relativi ai segni anamnestici, diagnostici e prognostici. L'individuazione di una corretta prognosi e le possibilità di guarigione ad essa collegate sono strettamente in relazione all'individuazione di precisi parametri temporali. Pertanto, in questo corpus testuale definito "testi di settore", si sono individuati due gruppi tematici: l'uno appartenente all'ambito della ricerca, l'altro della clinica; infatti, le forme grafiche che compaiono maggiormente nella classe di alte frequenze nel vocabolario del corpus "testi di settore" sono "soggetti" e "pazienti". Si può dunque pensare che questi repertori rintraccino da un lato, gli oggetti delle ricerche scientifiche, dall'altro i "protagonisti" degli interventi terapeutici.

In conclusione, se viene confermato che la conoscenza scientifica pervade le descrizioni del senso comune, contribuendo a modificarle o a consolidarle, emerge altresì con forza come le rappresentazioni sociali (il senso comune) influiscano in maniera altrettanto determinante sui processi di costruzione del senso scientifico.

Bibliografia

Berger P. e Luckmann T. (1966). *The social construction of reality*. Garden City.

Bolasco S. (1999). *Analisi multimediale dei dati*. Carocci.

Canguilheim G. (1966). *Le normal et le pathologique*. Presses Universitaries de France.

Foucault M. (1961). *Histoire de la folie à l'âge classique*. Gallimard.

Gergen J.K. (1985). Il movimento socio-costruzionista nella psicologia moderna. *American Psychologist*, vol. (4/3).

Losito G. (1993). *L'analisi del contenuto nella ricerca sociale*. Franco Angeli.

Fiora E., Pedrabissi L. e Salvini A. (1988). *Pluralismo teorico e pragmatismo conoscitivo in psicologia della personalità*. Giuffrè.

Salvini A. (1998). *Argomenti di psicologia clinica*. Upsel Domeneghini Editore.

Turchi G. P., Di Maso A., Mussoni A. e Salvini A. (2000). Identità as social and self narration: an approach to clinical psychology. In Doré F. (Ed.), *International Journal of Psychology*, vol. (35).

"Moved deeply I am"
Autistic language in texts produced with FC

Arjuna Tuzzi[1], Marisa Cemin, Marco Castagna

[1]Dept. of Historical and Political Studies – via del Santo, 28 – 35123 Padova – Italy
arjtuzzi@stat.unipd.it

Abstract

Using texts produced through Facilitated Communication (FC), this work is aimed at identifying the characteristic features of the language used by autistic subjects and understanding when these distinctive elements may distinguish it from the language of facilitators. Preliminary results shows that autistic subjects actually use a special style of writing; this finding supports the hypothesis that texts are the fruit of individual production of autistic subjects, not inevitably influenced by facilitators. This first work, based on a restricted sample which is not necessarily representative, is important because it has permitted to better specify criteria by which subjects, texts, analyses to carry out and software to be employed will be chosen in future studies. Such a protocol may be later applied to a broader interdisciplinary project involving linguistics, statistics, computer science, pedagogy, neurology, psychology, social sciences and ethical-philosophy. This new project, named "EASIEST" (Autistic Expression: Interdisciplinary Study with Statistic-Textual data processing) is based on the idea that in the specific case of autism, each form of communication represents a resource to be evaluated even if this seems atypical and controversial.

Keywords: textual-approach, software Taltac, vocabulary comparison, characteristic textual units.

Acknowledgments:
We wish to thank people in charge of the institutes for Facilitated Communication training who have supplied us with data for this work: Patrizia Cadei, Vittoria Cristoferi and Sergio Vitali; and to thank also Lorenzo Bernardi, Sergio Bolasco and Patrizio Tressoldi for their useful insights. We are also very grateful to the (autistic) authors of the texts and their facilitators.

1. Introduction

Autism is a serious illness that affects the development of human neurological organisation. Nowadays it affects in its most serious form one children out of 2,000. As its origin is unknown, there is no resolutive remedy. The biggest problem for people affected by autism and their families is represented by difficulties to respond to other people and to interact with the surrounding world (autistics can be either totally dumb or speak with whimsical and strange or repetitive language expressions). It is therefore necessary to find effective ways of communication (including those alternative to spoken language) in order for people affected by autism to open up new spaces of exchange with their families, to start suitable educational programmes and to find possible solutions to relieve their troubles.

Facilitated Communication (FC) is a technique (Biklen, 1999; Crossley and McDonald, 1990) that aims at creating a situation to communicate between two (or more) people. After a specific training, FC can offer the possibility to interact through the use of a hardware (i.e. a personal computer) which helps the writing of messages and through a human support (the facilitator) who gives physical and emotional sustainment to the autistic subject. During a FC session the autistic subject and the facilitator are placed side by side in front of the PC and

chat using video and keyboard. The facilitator usually stimulate the conversation digiting a question and giving the autistic subject the possibility to react (for instance by digiting the answer). During the first sessions gestures of the autistic subject are helped by the facilitator's hand placed on the subject's wrist. The contact gradually moves to the elbow, arm, shoulder, back and knee (according to the specific case) in order to obtain the maximum autonomy of the subject. However there is no general agreement about FC reliability both as an educational method and a way of communication: in particular, those who affirm autistic cognitive incompetence are doubtful about the autenticity of their texts, considering them the fruit of the influence (and great sleight) of facilitators (Green, 1994; Jakobson *et al.*, 1995). Moreover the validation method is still a matter of dispute among researchers all over the world. The idea of a survey on texts produced with FC from a statistic-linguistic point of view (Niemi and Kärnä-Lin, 2002; Zanobini and Scopesi, 2001) enters the above mentioned debate, since researchers try to find out useful methods to assess if (and how) texts written by autistic subjects are different in style and lexical choices from texts written by facilitators.

The present study uses texts written during FC sessions by subjects with clinical diagnosis of autistic disorder (according to DSM-IV). Three Italian institutes have provided us with texts: the "Center of Study on Facilitated Communication" (Chiavari, Genova), directed by Patrizia Cadei; the "Experimental Center for Development and Communication Disorders" (Padua), directed by Vittoria Cristoferi Realdon and the Social Cooperative "Intervento" (Mestre, Venice), directed by Sergio Vitali. The written works of 11 people aged from 9 to 22 have been collected: everybody had reached the elbow facilitation level and had been writing on a PC for at least a year with expert/habitual facilitators. This implies that all the texts we have at disposal have been written by autistic subjects who have been writing for a long time, together with expert/habitual facilitators, meaning a high fluency and a high level of autonomy.

2. Texts organisation and coding

The texts written with the FC technique can be different: a daily dialogue concerning the activities done during the day; a test on a school lesson; the composition of a poem; an educational speech; a free composition; information about personal facts, etc.

Capital letters are used to designate the part of the text written by the facilitator, whereas lower-case letters are used by the autistic subject in order to distinguish them clearly. To give an example:

COME HAI IMPARATO A LEGGERE E SCRIVERE?	HOW DID YOU LEARN TO READ AND WRITE?
a scuola guardando gli altri leggere, la cf è servita a farlo capire	at school looking at others reading, fc let it know
SEI STATO ALLA SCUOLA MATERNA? TI PIACEVA? COSA TI DISTURBAVA?	HAVE YOU BEEN AT PRIMARY SCHOOL? DO YOU LIKE IT? ANYTHONG TROUBLED YOU?
ma andiamo alla preistoria. per me è nebbia	we go back to prehistory. it's fog for me
C'E' QUALCOSA CHE VORRESTI DIRE AGLI INSEGNANTI E ALLE ALTRE PERSONE PRESENTI IN SALA?	WOULD YOU LIKE TO TELL SOMETHING TO THE TEACHERS AND THE OTHER PEOPLE IN THIS ROOM?
l'autismo è guardare voi che vivete, amate, lavorate e intanto il cielo è sempre più blu, notte.	autism is looking at you living, loving, working , meantime the sky is more and more blue, night
(P.C., anni 21)	(P.C., 21 years old)

Two separate corpora have been therefore created in order to understand if there is a specific character of the autistic language and if there are perceptible differences with the language of

facilitators: 1) the first comprising the texts of the facilitators (questions and comments which have appeared during the sessions); 2) the second concerning the texts written by the facilitated autistic subjects (answers, comments, poems, written compositions, etc.). Since the two corpora were limited in dimension, no further distinction has been made (age, facilitation level, place of the session, argument etc.) as the subparts could be of little relevance for the analysis.

The two corpora have been studied using both simple word-types (graphic forms) and complex textual units coding procedures (fig.1). In the first phase only simple word-types have been chosen (Lebart *et al.*, 1998). Step by step, a more complex recoding have been developed using an Italian common use multi-words list (phrase and polyrhematic recognition) and the extraction of repeated segments (multi-words and polyforms that make sense and other relevant sequences of words that gain or change meaning if considered as a block). All these operations have been done with the Taltac software (Bolasco *et al.*, 2000) and applying linguistic resources to process Italian language.

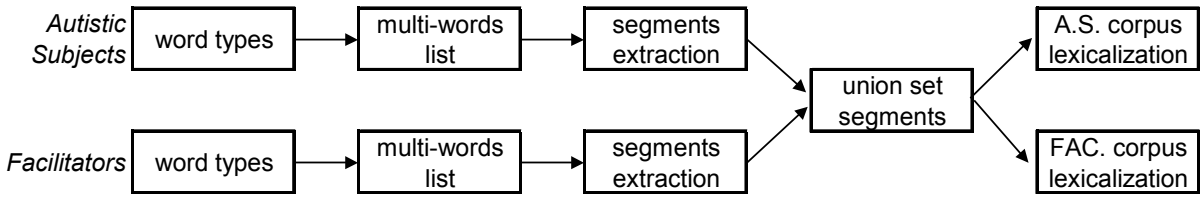


Figure 1. Coding procedure of the corpora into complex textual unit

3. Comparisons based on simple word-types vocabularies

3.1. *Autistic subjects*

The corpus of texts written by autistic subjects includes a total of N=24,499 word-tokens and $V(N)$=5,264 word-types. The Type-Token Ratio ($T.T.R.=V(N)/N$=21.5%) and the hapax percentage ($V_1/V(N)$=62.9%) point out the presence of a rich language; a result which can be partially explained through the limited dimension. The word-types vocabulary offers a very interesting description of the texts: at the top of the list which is ordered per decreasing frequency there are in fact some word-types showing a discourse which is mainly centered on self (‘mi / myself’, ‘io / I’, ‘sono / I am’, ‘ho / I have’, ‘me / me’, ‘voglio / I want’, ‘mio / my-male’, ‘mia / my-female’, ‘tu / you’, etc.). We can notice a massive use of negative forms (‘non / not’, ‘ma / but’, ‘no’) and of the word ‘molto / very’.

In order to read the vocabulary in a comparative perspective a reference lexicon taken from the Italian newspaper “La Repubblica” stored in the Taltac software has been used. It consists of a wide set of high frequency word-types (more than 50 occurences in ten years of the newspaper publications) and is a good source of written Italian language. Such reference lexicon is useful to extract the key words from the corpus calculating over/under-usage with respect to the frequency of the lexicon assumed as reference language. If only word-types that autistic texts and reference lexicon have in common are taken into account (4,417 word-types of the intersection set out to 5,264, equal to 84.9%) and indexes of over/under-usage are computed (Bolasco, 1999) we can notice the over-usage of some topic words such as ‘autismo / autism’ and ‘autistico / autistic-male’ and of words used for answers and questions such as the names (of the facilitators and of the subjects) together with other simple word-types such as ‘ciao / hi’, ‘sì / yes’, ‘ok’, ‘no / no’, etc. Moreover, there are elements that emphasize the per-

sonal nature of the exchange: the effort to improve the autistic condition is visible ('voglio / I want', 'devo / I must', 'tento / I try', 'riesco / I succeed', 'imparo / I learn', 'comunico / I communicate') together with the suffering ('aiutami / help me', 'rompo / I bother', 'resisto / I resist', 'mostrarmi / show me', 'deficiente / idiot') and the communication of feelings, wishes, thoughts ('sento / I feel', 'penso / I think', 'empatia', 'gioia / joy', 'pauroso / frightened'). On the contrary, if only word-types showing originality (793 word-types of the complementary set out of 5,264, equal to 15.1% of the vocabulary, that occur in the texts of autistic subjects and never in the reference lexicon) are considered, some obvious topic words appear such as 'autistica / autistic-fem.', 'autistici / autistic-plur.' and very unusual words like 'opto / I opt', 'tremo / I shiver', 'gioisco / I am glad', 'opero / I operate', 'controllarmi / control myself', 'limbico / limbic', 'navigo / I sail'.

3.2. *Facilitators*

The facilitators' corpus comprises a total of N=23,386 word-tokens and $V(N)$=3,714 word-types. The Type-Token Ratio ($T.T.R.=V(N)/N$=15.9%) and the hapax percentage ($V_1/V(N)$=57.4%) show that the language is in this case less rich than the one of the autistic subjects. This is a consideration based on a limited corpus that can nevertheless be compared with the other in terms of size and contents. The word-types vocabulary begins with 'che / what / which' followed by 'di / of' and 'ti / you': these are all devices that suggest the presence of a language based on questions ('che cosa...? / what...?', 'ti...? / do you…?', 'perché...? why…?'). In the high and medium frequencies areas of the vocabulary there is also the presence of many colloquial words such as 'vuoi / you want', 'fare / to do', 'oggi / today', 'anche / also / too', 'questo / this', 'quando / when', 'devi / you must', 'dire / say', etc. The comparison with the "La Repubblica" reference lexicon allows to observe that in the part in common (3,196 word-types of the intersection set out of 3,714, equal to 86.0%) there are some word-types which are over-used by facilitators and that refer again to the question wording style: 'ciao / hi', 'ti / you', 'hai / you have', 'vuoi / you want', 'scrivi / you write', 'vorresti / you would', 'stai / you are', 'ok', 'dimmi / tell me', 'senti / listen'. The result is not surprising because the occurrence of direct speech question/answer is not a characteristic feature of journalistic language, but on the other hand is very common in FC sessions. The analysis of word-types showing originality is probably more interesting (518 word-types of the complementary set out of 3,714, equal to 13.9% that occur in the texts of facilitators and never in the reference lexicon) because in this case the language of the educator is recognizable (FC is not only a way to communicate but also an educational method): 'intendi / you mean', 'definisci / you define', 'aiutarti / help you', 'riferisci / you refer', 'controllarti / control yourself', 'capirti / understand you', 'descrivi / you describe'. Unlike the case of autistic subjects, here non-common use elements have not been found.

3.3. *Autistic subjects versus facilitators*

Following the same criterion as the case of the reference lexicon, we extract word-types showing originality directly comparing the word-type vocabulary of the facilitators with the one of autistic subjects. The two vocabularies have in common 1,950 word-types. This intersection set represents 52.5% of the facilitators' vocabulary (1,950 word-types out of 3,714) and only 37.0% of the vocabulary of the autistic subjects (1,950 out of 5,264). Facilitators use 1,764 word-types never used by autistic subjects (equal to remaining 47.5% of the vocabulary), whereas autistic subjects put in their texts 3,315 word-types never used by facilitators (equal to 63.0%). This information (also represented in fig.2) suggests a greater wealth of words expressed by autistic subjects. However it is also interesting to take into consideration

the qualitative features of these words. The word-types showing originality of the facilitators (never used by the autistic subjects) always reflect an educative language based on questions: 'vorresti / you would', 'intendi / you mean', 'definisci / you define', 'insegnarti / teach you', 'controllarti / control yourself', 'scolastico / at school', 'bravissimo / well done', etc. On the contrary, word-types showing originality of autistic subjects (never used by the facilitators) refer to a wider world of language: 'potrei / I could', 'dolce / sweet', 'aiutami / help me', 'fede / faith', 'scusa / sorry', 'fortunato / lucky', 'anima / soul', 'mite / gentle', 'opto / I opt', 'vincere / to win', 'amica / friend-fem.', 'sicurezza / security', 'tristi / sad-plur', 'volo / I fly', 'feroce / cruel', 'ferito / wounded', 'cielo / sky', 'caos / chaos' , 'eticamente / etically', 'oltremodo / extremely', 'fresca / fresh-fem', 'fuoco / fire'.

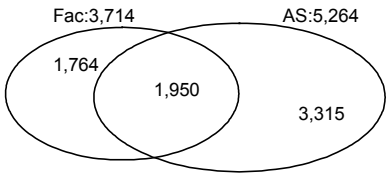


Figure 2. Comparison of the word-types vocabularies of facilitators and autistic subjects

It seems interesting to underline that each time the autistic subject uses an unknown (or not completely clear) for the facilitator word during a FC session, the facilitator immediately looks for an explanation about the use of the term (in the example given the names have been censured). Peculiar words of autistic subjects are therefore widely underestimated since during FC sessions some peculiar terms can be pronounced first by the facilitated autistic subject and then by the facilitator. This matter needs further investigation.

CIAO NOME_SOGGETTO_AUTISTICO
domani vieni in villa?
NO, DOMANI NON CI SARO'.
non buona cosa mio buon nome_facilitatore,
nome_soggetto_autistico corre seri rischi
IN CHE SENSO?
trappola mortale villa **adito** un'opportunità
CHE VUOL DIRE **ADITO**?
vuol dire grave trappola
PUOI ESSERE PIU' ESPLICITO?
essere più esplicito verrete voi traditi da me **itto**
ITTO?
sì
CHE COSA SIGNIFICA?
vuol dire nome_soggetto pazzo
(G., anni 23)

HI AUTISTIC_SUBJECT_NAME
Will you come to the villa tomorrow?
NO, TOMORROW I WILL NOT BE THERE
Not a good thing my dear facilitator_name,
autistic_subject_name is exposed to risk
WHAT DO YOU MEAN?
Mortal trap villa **adito** an opportunity
WHAT DOES **ADITO** MEAN?
It means serious trap
COULD YOU BE MORE EXPLICIT?
Being more explicit will you betrayed by me **itto**
ITTO?
yes
WHAT DOES IT MEAN?
It means autistic_subject_name crazy
(G., 23 years old)

4. Comparisons based on complex textual units vocabularies

4.1. *Autistic subjects*

As shown in text-figure 1, some intermediate phases for the analysis have been fixed in order to code the two corpora into new complex textual units. First of all we used a phrase and polyrhematic recognition based on an integrated Italian common multi-words list of the Taltac software; then we extracted repeated segments: multi-words and polyforms that are not contained in the Taltac lists and sequences of (less then seven) words repeated at least two times in the corpus. These non-hapax segments are 1,166 in total. This list shows the inci-

dence of segments comprising 'non / not' (148 segments) and 'molto / very' (33). There are a lot of emotional expressions addressed to their facilitators ('mio buon nome_facilitatore / my dear facilitator_name', 'furba nome_facilitatrice / she-cunning facilitator_name). Moreover, there are some segments lacking grammatical elements (especially conjunctions and auxiliaries).

Grammatical elements may also be placed in positions that do not respect the correct structure of sentences in the Italian language. To give some examples: 'penso proprio non lo so / I think definitely I don't know', 'molto voglio dirvi grazie / a lot I want to say you thanks', 'sono solo molto pieno di pesante disperazione / I am only full very of heavy desperation', 'grande sono diventato / grown I have become', 'molto bravo sono con mia resistenza / clever very I am against my resistance', 'molto sono arrabbiato con te / angry very I am with you', 'molto sono emozionato / moved deeply I am', 'questo voglio fare non / this I want to do not'. These segments are undeniably interesting on linguistic grounds; it is also quite easy to recognize the qualitative specificity of these expressions. However, the evaluation of this hypothetical "autistic style" goes beyond this work as it ought to be studied by linguistics.

According to a statistic method, the Morrone's IS index (Morrone, 1996; Bolasco *et al.*, 2000) has been applied together with a manual selection in order to reduce the 1,166 segments to 555 (segments that make sense without redundancy). This index is able to recognize a meaningful sequence of word-types comparing the frequency of the sequence as a block with the frequency of its word-types components. This new list has been used for the lexicalization of the corpus in order to obtain a recoding of the corpus in new complex textual units (word-types, multi-words, polyforms, polyrhematic, idioms, and segments together in the same vocabulary).

4.2. *Facilitators*

The same criterion has been followed for the corpus containing the facilitators' texts. Running the Taltac procedure of recognition of common use multi-word and extracting non-hapax segments of length below seven (a list of 1,405 multi-words reduced to 799 using Morrone's IS index and a manual selection) we have obtained a list for lexicalization. By means of a comparison between the two lists (facilitators and autistic subjects) some differences have appeared. Among the segments of the facilitators with high absolute IS index (long sequences extraction) there are common use expressions like: 'tu mi chiami / you call me', 'ti rendi conto / you realise', 'sicuramente non avrai problemi / surely you will not have any problem', 'hai lavorato bene / you have well worked', 'non hai risposto / you did not answer'. The list of segments of autistic subjects shows expressions like: 'nidifico uova in vista di rivelarmi / I'm nesting eggs in order to reveal myself', 'abbandonato hai questa sterile vita / you left this sterile life', 'sforzo non potrà limitarmi / will not strain be able to limit me', 'opulenza di eccessi / opulency of excesses'. A technical language appears in the segments of the facilitators with high relative IS index (short sequences extraction): 'comunicazione facilitata / facilitated communication', 'materie scolastiche / school subjects', 'chiesa romanica / Romanesque church', 'maestra nome_maestra / teacher teacher_name'. This language differs in qualitative terms from the corresponding expressions of the autistic subjects: 'cultura umanistica / humanism', 'completa autonomia / complete autonomy', 'era nera / dark age', 'buon nome_facilitatore / dear facilitator_name', 'oserei dire / I dare say'.

4.3. *Autistic subjects versus facilitators*

The study of the two lists of selected segments suggests that the two languages are different as shown by the dimension of the intersection set of the two lists: only 62 common segments appear (fig.3).

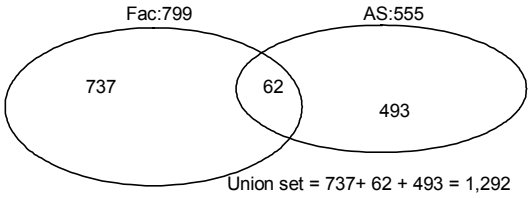


Figure 3. Comparison of the lists of segments

If the two lists are merged together, a union set of 1,292 segments are obtained. This list has been used for the lexicalization of both corpora in order to obtain a partition comparable textual units. In fact, this operation creates two new vocabularies of complex textual units which can be compared. This comparison confirms what other analysis have already suggested (fig. 4): the facilitators' vocabulary is only partly overlapping the autistic subjects' vocabulary. In addition to this, the forms contained in the non common parts are different in quality.

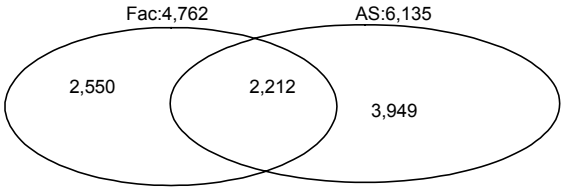


Figure 4. Comparison of the complex textual units vocabularies of facilitators and autistic subjects

In order to recognize the textual units present a great deal more (or a great deal less) in the facilitators' corpus than in the autistic subjects' one (elements useful to distinguish between the two groups) we can use the traditional 'characteristic textual units' method (Lebart *et al.*, 1998). This simple tool is based on the hypergeometric model (Lafon, 1980) and by means of a probability of over-usage it can detect which elements are used frequently by a group (as well as which elements tend to be used rarely by means of a probability of under-usage). The hypergeometric model is not sufficient to draw general conclusions about differences between the two corpora because the researcher has no indication of the distribution of these 'characteristic textual units' within the corpus. In fact a textual unit can be identified as specific of the corpus even if it occurs only in the text of one or two facilitators (or autistic subjects) rather than being representative of the group as a whole. A further test on dispersion (Tuzzi and Tweedie, 2002; Tuzzi, 2003) has been used to select only 'characteristic textual units' that are not confined to a subset of the texts. The following list shows the 'characteristic textual units' (ranked in decreasing frequencies order) which can be therefore considered suitable to describe the specificity of the corpus written by autistic subjects versus the facilitator's one.

Autistic subjects: 'io / I', 'ma / but', 'no', 'molto / very', 'me', 'ho / I have', 'mio / my-male', 'voglio / I want', 'grazie / thanks', 'vita / life', 'niente / nothing', 'senza / without', 'penso / I think', 'voi / you-plur.', 'mondo / world', 'dentro / inside', 'io sono / I am', 'aiuto / help', 'bisogno / need', 'sentire / to feel', 'amore / love', 'dire / to say', 'uomo / man', 'miei / my-plur.', 'deve / must', 'gioia / joy', 'mi sento / I feel', 'vivere / to live', 'io ho / I have', 'penso

che / I think that', 'pensieri / thoughts', 'dico / I say', 'desiderio / wish', 'mi piace / I like', 'riesco / I succeed', 'sento / I feel', 'non lo so / I do not know', 'non voglio / I do not want', 'autistico / autistic-male', 'non posso / I cannot', 'aiutami / help me', 'dolce / sweet', 'sole / sun', 'sentimento / feeling', 'fortunato / lucky', 'mio autismo / my autism', 'io voglio / I want', 'sono stanco / I am tired', 'molto bene / very well', 'dentro di me / inside me', 'io devo / I must', 'ma voglio' / 'but I want', 'come me / like me'…

Facilitators: 'perché / why', 'ti / you', 'cosa / what', 'come / how', 'se / if', 'te / you' 'sei / you are', 'anche / also / too', 'hai / you have', 'quando / when', 'oggi / today', 'chi / who', 'ok', 'allora / then', 'tuo / your-male', 'vuoi / you want', 'tua / your-female), 'ora / now', 'fatto / done', 'sai / you know', 'adesso / now', 'quale / which', 'quali / which', 'altro / other', 'capire / to understand', 'secondo te / in your opinion', 'a casa / at home', 'come stai / how are you', 'in classe / in the classroom', 'ti piace / do you like it', 'scrivi / you write', 'definisci / you define', 'cosa vuol dire / what does it mean', 'cosa intendi / what do you mean', 'cosa significa / what does it mean', 'cosa succede / what's happen', 'come ti senti / how do you feel', 'cosa vuoi dire / what do you mean', 'in che senso / in what sense', 'cosa vuoi fare / what would you like to do', 'cosa pensi / what are you thinking', 'controllarti / control yourself', 'una domanda / a question', 'devi andare in bagno / have you to go to the toilet', 'devi imparare / you must learn', 'scrivi qualcosa a piacere / write something at will'.

5. Conclusions

Despite the fact that the validity of the FC technique has not been discussed in this paper, it is nevertheless impossible to ignore the striking nature of the language produced by autistic subjects during the sessions: they write in a very special way with a sparing use of words; unusual terms are used, short phrases are preferred and they talk about emotions, intentions and feelings.

According to this, many issues need further investigation and bigger and more relevant corpora. Which kind of language is the one which is never orally expressed? If an autistic language exists, how can the specific features that distinguish it from the one used by facilitators be explained? However, the main priority of people applying FC is obviously to increase the quantity and improve the quality of autistic subjects' communication (Vidal *et al.*, 2002).

The results obtained should never be used to further isolate people affected by autism, considering them like strange speaking individuals with an impenetrable and enigmatic inner world.

References

Biklen D. (1999). *La comunicazione facilitata*. Omega edizioni.

Bolasco S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Bolasco S., Baiocchi F. and Morrone A. (2000). *TALTAC versione 1.0*. CISU.

Crossley R. and McDonald A. (1990). *Annie's Coming out*. Penguin Books.

Green G. (1994). Facilitated Communication: Mental Miracle or Sleight of hand. *Skeptic*, vol. (2/3): 68-76.

Guerra C., Cadei P. and Battistoni N. (Eds) (2002). *A voce alta*. Edizioni junior.

Jacobson J.W., Mulick J.A. and Schwartz A.A. (1995). A History of Facilitated Communication: Science, Pseudoscience, and Antiscience. *American Psycologist*, vol. (50/9): 750-765.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, vol. (1): 127-65.

Lebart L., Salem A. and Berry A. (1998). *Exploring Textual Data*. Kluwer Academic.

Morrone A. (1996). Temi generali e temi specifici dei programmi di governo attraverso le sequenze di discorso. In Villone M. and Zuliani A. (Eds), *L'attività dei governi della repubblica italiana (1947-1994)*. Il Mulino: 351-69.

Niemi J. and Karna-Lin E. (2002). Grammar and lessicon in facilitated communication: a linguistic authorship analysis of a finnish case. *Mental Retardation,* vol. (40/5): 347-357.

Tuzzi A. (2003). *L'analisi del contenuto.* Carocci.

Tuzzi A. and Tweedie F.J. (2000). The Best of Both Worlds: Combining MOCAR and MCDISP. In *Actes des JADT 2000*: 271-276.

Vidal J.M., Quris R., Morin A., Guillemot P., Huau V. and Coq J.M. (2002) L'analyse textuelle comme outil en sémiologie clinique: repères dans les expressions autistiques. In *Actes des JADT 2002*: 767-777.

Zanobini M. and Scopesi A. (2001). La comunicazione facilitata in un bambino autistico. *Psicologia clinica e dello sviluppo*, vol. (3): 395-420.

Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? L'exemple du projet PRINCIP

Mathieu Valette[1,3], Natalia Grabar[1,2]

[1]CRIM – INaLCO – 2 rue de Lille – 75343, Paris cedex 07 – France
[2]STIM/DIAM, DSI AP-HP – Paris 6 – 91 bd de l'Hôpital, 75634 Paris cedex 13 – France
[3]UMR 7114 MoDyCo – Paris 10 – 200 av. de la République, 92001 Nanterre Cedex – France
mathieu.valette@inalco.fr
ngr@biomath.jussieu.fr

Abstract

The authorities' pressing needs regarding web-users' protection against illegal or abusive content on the Net — racism, xenophobia, paedophilia — have implied setting aside conventional key-word-based filtering systems as well as black lists, given their lack of efficiency and the need for frequent updating. The purpose of PRINCIP, the multilingual platform for filtering racist and revisionist pages on the web is to implement a global, multi-criteria differential semantic analysis of web pages based on textual statistics, phrase extraction, and the theoretical proposals of François Rastier's interpretative semantics. In this paper, we present the results obtained by combining the use of two different NLP tools: Lexter (Didier Bourigault) and Hyperbase (Étienne Brunet).

Résumé

La demande pressante des institutions en matière de protection des usagers contre les contenus illicites ou préjudiciables sur Internet (racisme, xénophobie, pédophilie) invite à dépasser les systèmes de filtrage automatique conventionnels basés sur des listes de mots-clés ou des annuaires d'adresses préétablies, peu efficaces et exigeant de fréquentes mises à jour. L'objectif de la plate-forme multilingue de détection de pages web racistes et révisionnistes PRINCIP est de mettre en œuvre une analyse sémantique globale, multi-critères, et différentielle des documents reposant à la fois sur les statistiques textuelles, l'extraction de syntagmes, et les propositions théoriques de la sémantique interprétative de François Rastier. Nous présentons ici les résultats obtenus dans cette optique en combinant l'utilisation de deux outils distincts, Lexter (Didier Bourigault) et Hyperbase (Étienne Brunet).

Mots-clés. Lexter, extraction de syntagmes, Hyperbase, statistique textuelle, sémantique interprétative, détection, filtrage, classification automatique, texte idéologique.

1. Problématique

PRINCIP (Plate-forme pour la Recherche, l'Identification et la Neutralisation des Contenus Illégaux et Préjudiciables sur l'Internet) est un projet de détection automatique des pages web racistes et xénophobes développé conjointement par plusieurs laboratoires de recherche européens[1]. Il repose sur une critique des systèmes de filtrage fondés sur des listes de mots-clés

[1] Financé par la Commission Européenne, dans le cadre du *Safer Internet Action Plan*, les partenaires linguistique du projet PRINCIP sont le Centre de Recherche en Ingénierie Multilingue (CRIM) de l'Institut National des Langues et Civilisations Orientales de Paris (INaLCO), l'Institut für Germanistik de l'université Otto-von-Guericke à Magdebourg, la School of Applied Language and Intercultural Studies (SALIS) de la Dublin City University (DCU).

qui témoignent d'une approche naïve du texte raciste, en suggérant qu'il y a des mots racistes et des mots qui ne le sont pas, sans considération pour leur mise en texte (ou condition d'énonciation). Autrement dit, ces systèmes reposent sur un préjugé ontologique discutable, comme si le racisme était une langue de spécialité avec une terminologie stable et univoque.

Or, en tant qu'expression d'une opinion, le racisme n'est pas un discours référentiel, mais il relève de la rhétorique. Par ailleurs, sa caractérisation et sa détection impliquent la prise en compte de l'*intertextualité* inhérente au web, manifestée, dans le cas présent, par la présence sur le réseau de sites *sur* le racisme, c'est-à-dire antiracistes, qui partagent avec les textes racistes une part non négligeable de leur vocabulaire. Enfin, pour des raisons qui tiennent à sa délictuosité, du moins au regard du droit français, le texte raciste fait la part belle à l'euphémisme.

Ainsi, « *bougnoule* », dans une perspective référentielle, c'est-à-dire détaché de ses conditions d'énonciation, est considéré par les locuteurs français comme un mot raciste ; mais il n'apparaît que de façon très marginale dans le vocabulaire raciste. Plus encore, l'analyse de nos différents corpus montre qu'il est trois fois plus fréquent sur les sites antiracistes que sur les sites racistes. Les auteurs racistes, dans la plupart des cas, préféreront gommer tout particularisme lexical et parler d'un « *jeune des cités* », ou d'un « *jeune des quartiers* », voire, simplement d'un « *jeune* ». Dès lors, les traits sémantiques caractéristiques du texte raciste se situeront en-deçà, ou au-delà du dictionnaire des mots-clés tels que « *bougnoule* ».

Pour caractériser les textes racistes, le recours aux statistiques s'est rapidement imposé comme cadre d'investigation. Mais quelles statistiques ? Une récente étude (Vinot *et al.*, 2003), menée en collaboration avec l'École Nationales Supérieure des Télécommunications (ENST, Paris), a permis d'évaluer les performance des algorithmes de classification automatique utilisés notamment dans les systèmes de filtrage du courrier non sollicité, ou *spam*. Ces algorithmes, qui n'ont évidemment pas de préjugés référentialistes, fonctionnent sur un mode contrastif, à partir de deux sous-corpus catégorisés (raciste et antiraciste) : en bref, il s'agit soit de calculer la distance euclidienne qui sépare la représentation vectorielle d'un document (tf*idf) des autres documents du corpus (algorithme k-PPV) ou d'une classe de documents (algorithme Rocchio), soit de distribuer tout nouveau document de part et d'autre d'un hyperplan séparant les données des deux sous-corpus (algorithme SVM).

Ces algorithmes ont présenté de bons résultats en ce qui concerne la classification des documents domiciliés sur des sites racistes dédiés, parce que ce sont les « signatures lexicales » de ces sites qui ont été discriminantes (sommaires, slogan, etc.). Mais ils restent peu efficaces lorsqu'il s'agit de détecter un alinéa ou une incise raciste dans un texte qui ne l'est pas dans son entier, ou lorsque le document est isolé (page « perso »).

2. Objectifs

La méthodologie mise en place dans le cadre du projet PRINCIP tient compte à la fois de l'inadéquation des approches référentialistes et des limites des approches statistiques, ou, plus positivement, il s'agit d'en combiner les avantages de façon à en neutraliser les défauts.

2.1. *Lexies racistes et mise en texte*

Si l'on peut difficilement parler d'Internet comme d'un corpus (Rastier, sous presse), on peut néanmoins parler de son *intertextualité* massive. Dans le cadre de notre problématique, celle-ci prend naissance dans la dialectique qui oppose les auteurs antiracistes aux auteurs racistes. La rhétorique antiraciste consiste en effet à déconstruire l'argumentation des textes racistes.

Une large place est allouée aux citations (que ce soit des lexies, des phrases, des paragraphes). Les lexies les plus stables et les plus ancrées dans le vocabulaire des auteurs racistes, c'est-à-dire celles qui feraient de bons candidats *a priori* à la constitution d'une liste de mots-clés, sont celles dont des auteurs antiracistes vont faire un usage critique privilégié parce qu'elles sont facilement identifiables (voir l'exemple ci-dessus, « *bougnoule* », ou le vocabulaire de l'extrême droite). Parallèlement, les auteurs racistes s'approprient certaines lexies antiracistes notoires. Par exemple « *pote* », emblème lexicale de l'association SOS-Racisme, s'il n'est plus guère utilisé par celle-ci que dans des lexies composées figées (ex. les associations de quartiers « *les maisons des potes* »), est remotivé par les auteurs racistes et utilisé à des fins euphémiques.

2.2. *Critères de collection, critères différentiels*

Finalement, le matériel conceptuel raciste constitue un point d'accès à la problématique du racisme. Il permet de collecter *tous* les documents relevant de cette problématique, qu'ils soient racistes ou antiracistes. On appellera ce matériel conceptuel les *critères de collection*. Ce sont les conditions de leurs actualisations qui seront discriminantes et permettront de distinguer les documents racistes des documents sur le racisme. Les critères discriminants qui permettent de contraster les deux sous-corpus sont appelés les *critères différentiels*.

En termes d'analyse de corpus, un nombre non négligeable de critères de collection ont été obtenus à l'aide du logiciel Lexter (Bourigault, 1994) — un analyseur syntaxique superficiel réalisé pour le compte d'EDF, originellement dédié à l'extraction de syntagmes (nominaux et adjectivaux) à partir de corpus spécialisés, dans une perspective d'acquisition terminologique. Les critères différentiels, purement statistiques, ont été obtenus à l'aide du logiciel Hyperbase conçu par Étienne Brunet[2] — un outil de lexicométrie et d'analyse documentaire. Prévu initialement pour traiter des textes littéraires (le dictionnaire de référence utilisé est le TLF issu de la base de données textuelles Frantext de l'INaLF, mais certaines fonctions s'affranchissent de cet héritage), Hyperbase est un outil d'analyse statistique accompagné d'un concordancier.

Ces deux logiciels ont été détournés de leur fonction première : d'une part, le racisme, comme nous l'avons vu, n'est pas une langue de spécialité et l'idée d'une terminologie raciste apparaît complètement infondée ; d'autre part, les textes racistes ne sont pas des textes littéraires mais relèvent, dans l'immense majorité des cas, du discours politique ou journalistique.

3. La constitution des corpus

L'outil final PRINCIP de détection de contenus illicites est destiné à traiter directement des documents collectés sur le net. Les corpus de travail ont donc également été construits à partir de données existant sur le Web. Nous avons utilisé les moteurs de recherche généraux que nous interrogions avec des mots clés « sensibles » *(*« *bande allogène* », « *gangs ethniques* », « *racaille black* », « *Sieg Heil* », etc.). La constitution du corpus a été faite en deux étapes : (i) collecte massive de documents ; (ii) catégorisation manuelle.

La collecte de documents fut réalisée de deux manières : interrogation manuelle et automatique de pages et de sites et leur rapatriement, à l'aide des outils (Grabar et Berland, 2001) et Unix Wget[3].

[2] UMR6039 *Bases, Corpus et Langage*, Université de Nice (http://ancilla.unice.fr).
[3] Sur la constitution du corpus PRINCIP, Cf. http://www-poleia.lip6.fr/~princip/PRINCIP-2119-D1.1.pdf.

Les données présentées ci-après ont été obtenues essentiellement à partir d'un corpus de sites contrasté raciste/antiraciste. Une typologie fine n'étant pas l'objet de cet article, on se contentera d'en décrire rapidement les grandes tendances :

Sites antiracistes

- explicitement dédiés à la lutte contre le racisme (*Dire et Faire contre le racisme*, *Jeune Contre le racisme en Europe*, *Antisémitisme Info*, *SOS-Racisme*, etc.) ;
- dédiés à des causes qui confinent à l'antiracisme : contre l'extrême-droite (*Plus jamais ça*, *Ras l'front*) ou pour la défense des migrants (*Groupe d'information et de soutien des Immigrés*, *Droit Humain*, *Hommes et Migrations*, etc.).

Sites racistes

- agressifs et haineux, explicitement « racialiste » (*AIPJ*, *CPIAJ*, *SOS-Racaille*, *NaziLauck*, *OdinsRage*, *Racist National Library*, etc.) ;
- plus policés, ciblant parfois selon des critères socioculturels plutôt que raciaux ou ethniques (*Unité radicale*, *les Identitaires*, *le Libre journal de la France courtois*, etc.).

4. Typologie des critères et technique d'obtention

Lexter (Bourigault, 1994) est un outil du Traitement Automatique des Langues Naturelles (TALN), dédié à l'analyse syntaxique superficielle des documents provenant d'un domaine de spécialité. Son but premier consiste donc en constitution de ressources terminologiques de ce domaine. Lexter traite des documents déjà étiquetés syntaxiquement dans lesquels il cherche les séquences syntaxiques (syntagmes nominaux et adjectivaux) aptes à représenter les *candidats termes* du domaine.

Lexter fonctionne en deux étapes :

- Recherche des frontières syntaxiques entre les candidats termes (verbes conjugués, pronoms, conjonctions, etc.). Les mots qui appartiennent à ces catégories ne peuvent pas faire partie des candidats termes et sont donc ignorés dans la suite des traitements.
- À l'intérieur des segments des phrases délimités, recherche des patrons prédéterminés qui peuvent correspondre à des *candidats termes*. Les suites de type *Nom Adj*, *Nom Prep Nom*, *Nom Prep VerbeInfinitif* constituent les patrons potentiels des termes.

Hyperbase est un logiciel d'analyse lexicométrique proposant toute une palette d'outils documentaires (deux concordanciers, dont l'un fournit les alignements et l'autre les contextes d'attestation) et statistiques (combinant des fonctions d'analyse structurelle, i.e. des informations sur la distribution du vocabulaire dans un corpus : spécificités, distances, étendue, richesse, accroissement, hautes fréquences, évolution etc. ; et des fonctions statistiques sur listes de formes constituées manuellement (analyses arborées, analyses factorielles).

4.1. *Des candidats termes au « candidats lexies »*

Les candidats termes sont des lemmes dont on a perdu le lien au contexte. Autrement dit, c'est la substance lexicale du texte raciste. On a compté pas moins de 62 000 items pour notre seul corpus raciste. Ces données brutes de Lexter sont des lexèmes (ou des lexies simples : verbes, noms, adjectifs, adverbes) et des lexies composées (syntagmes nominaux, syntagmes adjectivaux). Toutes les sorties sont lemmatisées. Comme pour les termes de métiers, la tâche consiste alors à expertiser ces lexies de façon à ne conserver que les bons candidats.

De nombreux « candidats lexies » ont été validés à partir des sorties de Lexter. Prenons l'exemple des syntagmes nominaux, c'est-à-dire des sorties comprenant 2 lemmes, généralement un nom et un adjectif. À partir des sorties de Lexter, on peut distinguer grossièrement trois ensembles de candidats lexies (composées) :

4.1.1. Les hautes fréquences, qui relèvent de la *signature lexicale* des sites, et qui le plus souvent proviennent du péritexte de la page web[4]. Par exemple « *vrai visage* », « *république islamique* », « *égorgement de femme et d'enfant* », etc. Un coup d'œil dans un concordancier suffit à reconnaître ces lexies récurrentes, aux contextes chaque fois identiques. Si elles constituent un matériau valable pour notre tâche de détection, du point de vue de la caractérisation du texte raciste, elles présentent l'inconvénient de faire écran aux données moins spécifiques à un ou quelques sites particuliers. Ces limites sont à rapprocher de celles observées à propos des algorithmes de classification (cf. § 1)

4.1.2. Les fréquences moyennes, sans doute les plus intéressantes, car c'est parmi elles que ce cachent les véritables lexies racistes, susceptibles d'être présentes sporadiquement mais de façon quantitativement remarquable. Par exemple, « *crime raciste* », « *islamiquement correct* », « *peuple français* », etc. Ces candidats lexies sont ensuite étudiés en contexte au moyen d'Hyperbase de façon à évaluer leur qualité en tant que lexies racistes d'une part, et en tant que lexies racistes susceptibles d'être actualisées dans un texte antiraciste d'autre part.

4.1.3. Enfin, les fréquences basses, plus difficiles à traiter, participent davantage à la dynamique investigatrice qu'à l'extraction de candidats lexies. Par exemple, elles permettent, par regroupement et recoupement d'informations, de mesurer le point d'un adjectif particulier, et la variabilité de ses instanciations, ou d'observer une construction néologique récurrente, etc. (par exemple, les différentes qualification de « *terroriste* » : « *terroriste beur* », « *terroriste musulman* », « *terroriste tchétchène* ».

4.2. *Actualisation des candidats lexies*

Les lexies racistes, comme nous l'avons dit précédemment, sont tout aussi bien actualisées dans des textes racistes qu'antiracistes. Une fois que des candidats lexies ont été retenus, il s'agit donc d'étudier leur actualisation, c'est-à-dire leur comportement en contexte et leur poids statistique. Pour ce faire, nous avons recouru à Hyperbase en particulier pour les lexies simples. Les lexies composées sont moins aisées à traiter au moyen de ce logiciel car d'une part, les fonctions statistiques leur sont inaccessibles et d'autre part, il n'est pas possible de lemmatiser un syntagme entier. Pour quelques calculs statistiques sur lexies composées, nous avons donc utilisé des programmes en Java composés à cet effet.

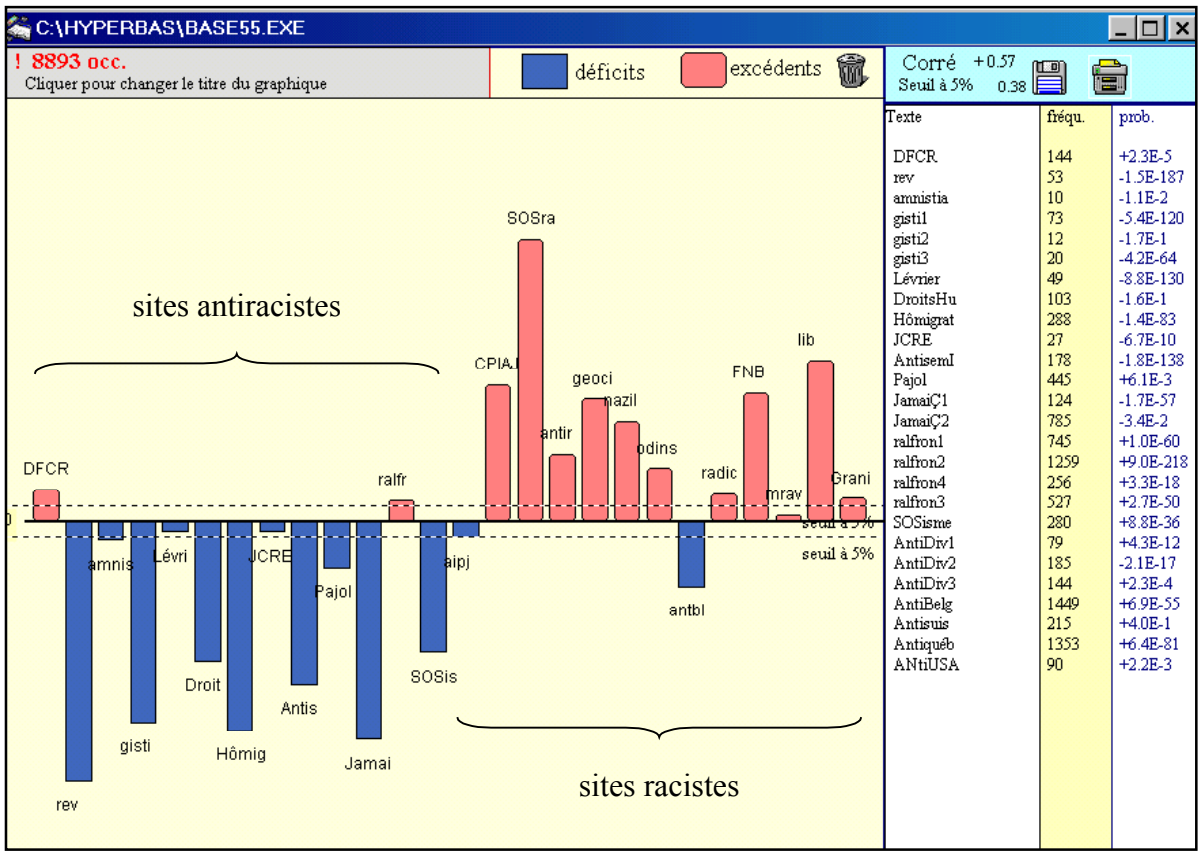
Pour illustrer le rôle des statistiques textuelles dans PRINCIP, reprenons un exemple de lexie raciste mentionnée ci-dessus : « *crime raciste* ». D'aucun n'auront pas manqué de trouver paradoxale cette lexie qui semble *a priori* (i.e. d'un point de vue référentiel) antiraciste. Certes, le lemme « *crime raciste* » est un peu plus fréquent dans les textes racistes que dans les textes antiracistes (55 % des occurrences apparaissent dans les textes racistes, contre 45 %

[4] La page web structurée en HTML, quel qu'en soit le contenu, est soumise à des contraintes intertextuelles fortes qui déterminent la forme du document et les formes du texte. Autrement dit, une page web, même vide, présente déjà un fond commun à toutes les pages du sites, que ce soit au niveau des étiquettes HTML elles-mêmes (structuration de la page, métadonnées) que des formes (lexicales) affichées à l'écran (par exemple le péritexte). L'ensemble constitue la *signature sémiotique* du site. Ainsi, Sur un corpus comprenant la totalité des pages d'un site raciste donné (www.sos-racaille.org, aujourd'hui fermé), nous avons mesuré que sur le texte seul, en moyenne 24,75 % des occurrences de formes appartenaient à ces informations péritextuelles communes à toutes les pages du site. Sur la source HTML, étiquettes et péritexte confondus, ce pourcentage atteint 47,45 %.

dans les textes antiracistes), mais cette proportion, calculée sur un nombre d'occurrences restreint, est assez peu pertinente. En revanche, sitôt que l'on contextualise un peu ce résultat, par exemple — c'est un minimum souvent discuté[5] — en distinguant la forme plurielle de la forme singulière, l'écart statistique se creuse : si la répartition de la première évolue peu, celle de la forme singulière stricte s'avère sensiblement plus raciste qu'antiraciste (66,17 % contre 33,83 %).

Il ne s'agit pas là d'un cas isolé mais d'une tendance lourde : le singulier est privilégié par les racistes quand, pour un même lemme, les antiracistes adopteront plus volontiers le pluriel. Ainsi, « *la femme* » vs. « *les femmes* », « *le peuple* » vs. « *les peuples* », etc. Un des exemples les plus marquants est « *droite nationale* », version auto-référentielle de ce qu'il est d'usage d'appeler l'extrême droite : 80 % des occurrences de la forme singulière sont racistes quand 89 % des occurrences de la forme plurielle — sensiblement plus rare en valeur absolue — sont antiracistes.

Pour opposer l'actualisation raciste et l'actualisation antiraciste, nous avons visé tout les paliers de la complexité textuelle, en nous focalisant au maximum sur les items qui ne relèvent pas ontologiquement des « domaines » du racisme. La liste des critères différentiels ci-dessous est évidemment loin d'être exhaustive, elle est indicative :



[5] Cf. récemment Brunet 2000.

4.2.1. *Les symboles*

Les points d'exclamation et les points de suspension sont plus fréquents chez les racistes ; à l'inverse, les antiracistes font un usage légèrement plus important des points-virgules et des parenthèses. L'histogramme ci-dessus (capture d'écran d'Hyperbase), illustre ce contraste avec les point d'exclamation.

Ce graphique illustre l'écart observé entre la fréquence du point d'exclamation et la fréquence théorique attendue[6] (le calcul pondère cet écart selon la formule de l'écart réduit). Sur le côté gauche de l'histogramme, on a sélectionné 13 textes antiracistes, et sur le côté droit, 13 sites racistes.

4.2.2. *Les morphèmes*

Parmi les morphèmes, on distinguera :

(i) les morphèmes rares, mais très discriminants, comme, par exemple, certains suffixes dépréciatifs ou utilisés de façon dépréciative (*-ouille*, dans « *démocrassouille* », *-âtre*, dans « *sémitolâtre* », *-mane*, ou *-manie*, dans « *israëlomane* »). Nous avons alors mesuré leur pondération en termes de précision et de rappel, par exemple :

	Rappel	*précision*
-ethn- (*ethnie*, *ethnique*, etc.)	25,60	59,23
-mafi(a)- (*mafia*, *mafieux*, etc.)	5,61	61,46
-ouill- (*magouille*, *fripouille*, etc.)	6,09	70,68
-man- (*israëlomane*, etc.)	23,65	68,37
-phil- (*crouillophile*, *philomarxiste*, etc.)	20,00	57,76

Rappel et précision racistes de quelques morphèmes à partir du corpus de test.

(ii) les morphèmes fréquents, tant dans le sous-corpus antiraciste que dans le sous-corpus raciste, mais avec un écart significatif. C'est notamment le cas de certains grammèmes. Par exemple, les prépositions *à* et *de* sont plus courantes dans les textes antiracistes, tandis les adjectifs possessifs relèvent du discours raciste. Ces résultats sont corroborés par un calcul effectué sur une base Hyperbase étiquetée à l'aide de Cordial.

4.2.3. *Les parties du discours*

Le résultat emblématique de l'étiquetage de notre corpus, sur lequel il y aurait lieu de gloser longuement, concerne la distribution des noms et des verbes : les antiracistes privilégient globalement les premiers, les racistes les seconds. Ce résultat est confirmé par d'autres informations telles que la longueur moyenne des mots. Les mots de l'antiracisme sont plus long en raison de la composition des substantifs (le suffixes–*ation*, plus fréquent, en est un bon indice). On pourrait avancer que le texte antiraciste est réflexif et objectif tandis que le texte raciste relève du passage à l'acte, du procès. Fidèle compagnon du verbe, l'adverbe est également très présent dans les textes racistes.

Si l'on s'intéresse aux étiquettes sémantiques proposées par Cordial, on remarque que les antiracistes manipulent davantage de noms abstraits que les racistes, et que ces derniers, à

[6] Fréquence théorique d'un mot dans un texte = fréquence du mot dans le corpus pondérée par la probabilité *p* ou part du texte dans le corpus.

l'inverse, privilégient les noms liés au corps humain et au règne animal. Cette observation est loin d'être anodine : comparer l'autre à un animal est un trait sémantique fondamental du discours raciste, où autrui est considéré comme un sous-homme, faisant partie d'une sous-race, ou d'une sous-espèce humaine.

Cette observation nous amène à nous intéresser aux lexies du racisme, c'est-à-dire, aux lexies qui constituent le vocabulaire raciste, mais qui ne sont pas, en elles-mêmes, racistes. Celles-ci peuvent être :

– des lexies simples : noms (« *chaos* », « *milliard* », « *sang* », « *destin* », « *larmes* », « *bombes* », etc.) ; verbes (« *falloir* », « *mourir* », « *dévaster* », « *massacrer* », etc.) ; adverbes (« *très* », « *jamais* », « *partout* », « *prétendument* », « *simplement* », « *vraiment* », « *naturellement* ») ; adjectifs (« *grand* », « *immense* ») ; mais aussi des dates (les dates récentes sont actualisées dans les textes racistes, les plus anciennes dans les textes antiracistes), etc.

– des lexies composées : « *coup de couteau* », « *aux mains de* », « *au cœur de* », « *en prise avec* », « *se mettre à genoux* », etc. et des lexies complexes : slogan (« *français d'abord* », « *ni raciste, ni xénophobe* »), entités nommées (« *Front uni anti-système* »), etc.

4.2.4. *Les thèmes sémantiques*

La fonction « thème » d'Hyperbase (mesure de l'attraction qu'un ou plusieurs mots pôles exercent sur leurs cooccurrents dans un corpus donné, à partir d'un test d'écart réduit) nous a permis de discerner des unités sémantiques complexes (isotopies sémantiques et molécules sémiques) dont le raffinement (éventuelle lemmatisation, extraction des radicaux) a été réalisé manuellement ensuite.

Citons par exemple l'isotopie 'animalité' dont il a été question ci-dessus. Nous avons isolé un certain nombre d'items (ici lemmatisés) qui relèvent de cette isotopie : « *femelle* », « *mâle* », « *bipède* », « *macaque* », « *bâtard* », « *chien* », « *rat* », « *cafard* », « *cloporte* », « *ramper* », « *peste* », « *choléra* », « *vermine* », « *proliférer* », « *grouiller* », « *puer* », etc. Par exemple :

> Sans doute attirés par l'**odeur infecte** des amoncellements de déchets locaux, une bande de **rats** d'Iran serait venue faire bombance au Liban (ah, l'état Liban !). Les **animaux** ont été signalés au nord de la frontière. En attendant, ils manifesteraient clairement l'intention de **nuire**. L'embêtant, c'est que maintenant il va falloir les détruire vite et bien jusqu'au dernier. Bref, depuis (plus de vingt ans) que le Shah n'est plus là, les **rats** dansent (site CPIAJ).

Pour la constitution du thème de la « tournante » (viol collectif), qui offre un excellent rappel dans les textes du genre fait divers, on peut étudier à titre d'exemple le thème suivant :

Écart	Corpus	Extrait	Mot	Écart	Corpus	Extrait	Mot
77.30	235	27	**viol**	25.09	76	5	**violée**
62.25	18	6	eiders	24.87	28	3	sabrina
60.95	207	20	**collectif**	23.83	54	4	vicieux
47.39	55	8	**tournante**	23.83	54	4	possession
32.62	29	4	**livrent**	23.83	54	4	enjeu
28.47	38	4	requins	22.59	60	4	**onze**
26.88	24	3	**répété**	22.40	61	4	**cibles**
26.74	43	4	acquérir	21.07	419	10	**blanche**
25.23	296	10	**blanches**	17.35	57	3	**blacks**

Écart	Corpus	Extrait	Mot	Écart	Corpus	Extrait	Mot
17.06	104	4	présenté	8.17	894	6	souvent
16.92	164	5	**racial**	8.15	639	5	doute
16.24	343	7	acte	7.53	733	5	**racistes**
15.51	194	5	**viols**	6.89	563	4	**noirs**
15.39	127	4	affirmer	6.70	592	4	**ethnique**
14.12	150	4	rap	6.38	2239	8	**ans**
13.76	348	6	**cité**	6.05	702	4	mai
12.80	1495	12	**jeunes**	5.93	1088	5	paris
12.55	107	3	**collectifs**	5.80	444	3	prison
12.26	750	8	**femmes**	5.78	758	4	**victimes**
10.30	273	4	agression	5.07	935	4	aucun
9.45	926	7	**groupe**	4.68	634	3	avril
9.23	192	3	libres	4.61	1083	4	**raciste**
9.13	737	6	**blanc**	4.57	658	3	lorsque
8.34	613	5	**femme**	4.24	3890	8	était
8.20	239	3	instant	4.03	1926	5	**fois**

Thème sur la cooccurrence des formes « tournante » et « viol »
sur un corpus composé exclusivement de textes racistes (seuil minimal de 3)

Si l'on met de côté les éléments trop conjoncturels (par exemple *requins* et *vicieux* se rapportent à un fait divers particulier), on voit se dessiner un thème qu'il nous faut ensuite compléter en croisant les corpus, en analysant les contextes et en recomposant les lexies déstructurées par l'analyse des formes isolées. En résumé, on retiendra 3 catégories de traits sémantiques :

– **Lieux** : avec le sème générique /extérieur/ (l'environnement : « *banlieue* », « *quartier* », « *cité* », etc.) ou /intérieur/ (le lieu du crime : « *cave* », « *sous-sol* », « *parking* », « *chambre* », etc.) ;

– **Actant** : /masculin/ (le bourreau : « *garçon* », « *copain* », « *compère* », etc.) ou /féminin/ (la victime : « *jeune fille* », « *jeune femme* », etc ;

– **Action** : /viol/ (« *tournante* », « *violer* », « *pénétration* », etc.).

À ce thème général, qui correspond à un fait divers banal, si l'on peut dire, et que l'on rencontre dans la presse non raciste, d'autres éléments lexicaux s'ajoutent et font basculer le texte dans la catégorie des racistes : c'est le cas des qualificatifs des bourreaux : « *pote* » historiquement emprunté au discours antiraciste à des fins euphémiques, mais surtout « *mâle* » qui comprend le trait /non humain/ présent dans l'isotopie 'animalité'. Quant à la victime, elle est « *blanche* », « *française* », ou encore « *gauloise* ». Enfin, les informations quantitatives ne doivent pas être négligées, car elles produisent un *effet de réel* : nombre des violeurs, et âge des actants (« *ans* »).

Par exemple :

> Une **adolescente** de **14 ans** rentre de l'école, cartable au bras. Elle est capturée par 3 "**jeunes**" , de 14 , 15 et 17 ans. Ils l'entraînent dans une **cave**, la **violent** tour à tour. Par la porte restée entrouverte, d'autres **adolescents** et enfants regardent, tout simplement.

En résumé, l'actualisation des candidats lexies racistes extraits du corpus par Lexter est évaluée en fonction de critères lexicaux, grammaticaux ou autres (un travail sur les étiquettes

HTML a également été mené en parallèle), pourvu qu'ils soient ontologiquement distincts des candidats lexies.

5. Discussion

Les objectifs différents des deux outils (Lexter pour l'acquisition terminologique d'un domaine de spécialité, Hyperbase pour l'analyse lexicométrique des documents littéraires) et, par conséquent, les fonctionnements et les fonctionnalités également différents, conditionnent les résultats qui peuvent être acquis avec chacun de ces outils.

Tout d'abord, les outils repèrent les critères de niveaux différents. Lexter fonctionne au niveau des lexies simples et des lexèmes (« *islam* », « *national* ») et des lexies composées (« *invasion islamique* », « *identité nationale* »). Hyperbase présente la possibilité d'interroger les documents sur les lexies simples et composées, avec la réserve émise précédemment (cf. 4.2.) ou encore les sous-chaînes de caractères (qui peuvent être rapprochées des morphèmes). Lexter ignore les symboles et la ponctuation (parenthèses, points, virgules, points d'exclamation). Hyperbase, pour des raisons de structuration des données, ne traitent pas quelques rares symboles ($, &).

Lexter effectue une analyse systématique des documents et la recherche de séquences correspondant aux patrons terminologiques. Avec Hyperbase, si l'utilisateur peut utilement recourir aux fonctions d'analyse de la structure du corpus, la recherche de critères est davantage amorcée d'une manière intuitive et ciblée par l'utilisateur. Le comportement de chaque critère est donc analysé suite à une requête le concernant. De là provient l'opposition entre le fonctionnement onomasiologique de Lexter (guidé par les données des documents) et sémasiologique d'Hyperbase (guidé par les idées et les intuitions de l'utilisateur étant donné ses connaissances et attentes vis-à-vis du corpus des documents).

Une dernière différence apparaît du fait que Lexter effectue une méta-analyse, ancrée sur les catégories syntaxiques des mots des documents. Il ne prend pas en compte les formes lexicales, alors que l'analyse effectuée avec l'Hyperbase est ancrée sur des chaînes lexicales données.

6. Synthèse

Nous avons présenté dans ce papier les travaux autour de l'établissement d'une base de critères pour la détection de contenus préjudiciables et racistes sur l'Internet. Les critères purement lexicaux n'étant pas suffisants, nous les complétons avec des critères d'autres niveaux :

- les caractères (symboles, ponctuation) ;
- les sous-chaînes de caractères (pouvant être rapprochées des morphèmes) ;
- les chaînes de caractères (mots ou lexies simples) ;
- la collocation de chaînes de caractères (lexies composées ou complexes).

La découverte et le recrutement de ces critères sont effectués avec deux outils. Lexter permet d'acquérir systématiquement tous les critères qui se trouvent aux niveaux des chaînes de caractères et de collocations (lexies). Hyperbase permet d'interroger le corpus sur les chaînes de caractères, les collocations, les sous-chaînes de caractères, etc. (lexies, morphèmes, graphèmes) prédéterminés par l'utilisateur, ainsi que sur les étiquettes morpho-syntaxiques, dans une version aménagée du logiciel.

D'autres types d'indices (liens hypertexte, code HTML, etc.) ont fait l'objet d'études séparées, avec des outils développés à cet effet. La qualité (précision) et la couverture (rappel) des indices est en cours d'évaluation sur un corpus de test. Les indices obtenus seront ensuite intégrés dans un système multi-agents développé par le Laboratoire d'Informatique de Paris 6.

Remerciements

Nous remercions Didier Bourigault d'avoir mis à notre disposition Lexter, Etienne Brunet d'avoir procédé à l'étiquetage de notre corpus pour Hyperbase, Monique Slodzian et François Rastier pour les discussions tout au cours de ce travail.

Bibliographie

Bourigault D. (1994). *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*. Thèse en informatique linguistique, École des hautes Études en Sciences Sociales, Paris.

Brunet Ét. (2000). Qui lemmatise dilemme attise. *Scolia, 11ème rencontres linguistiques en pays rhénan*, vol. (13) : 7-32.

Grabar N. et Berland S. (2001). Construire un corpus web pour l'acquisition terminologique. *Terminologie et intelligence artificiell*e : 44-54.

Malrieu D. et Rastier Fr. (2001). Genres et variations morphosyntaxiques.In Daille B. et Romary L. (Eds), *Linguistique de corpus, TAL*, vol. (42/2) : 548-577.

Muller C. (1964). *Essai de statistique lexicale. "L'illusion comique" de Pierre Corneille.* Klincksieck.

Rastier Fr., Cavazza M. et Abeillé A. (1994). *Sémantique pour l'analyse : de la linguistique à l'informatique*. Masson.

Rastier Fr. (sous presse). Enjeux épistémologiques de la linguistique de corpus. In Williams G. (Ed.), *Actes des deuxièmes Journées de linguistique de corpus de Lorient*. Presses Universitaires de Rennes.

Valette M. (2003). Détection et interprétation automatique de contenus illicites et préjudiciables sur internet : le projet PRINCIP. In Rastier F. (sous la direction de), *Texto ! Sémantique des textes* (http://www.revue-texto.net).

Vinot R., Grabar N. et Valette M. (2003). Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet. In *Actes de TALN 2003* : 257-284.

Analyse d'hyperliens en vue d'une meilleure description des profils

Valéry Vandaele, Pascal Francq, Alain Delchambre

CAD/CAM – Université Libre de Bruxelles – Av. F.D. Roosevelt, 50 – CP 165/14 1050 Bruxelles – Belgique

{vavdaele, pfrancq, adelch}@ulb.ac.be

Abstract

The objective of this paper is to discuss the integration of various algorithms of link analysis into the GALILEI project. In the frame of this projet aiming at clustering users into virtual communities according to their centers of interests, one notices indeed that the system converges more rapidly when a document is consulted by more than one user. Knowing that the methods of link analysis could emphasize the interesting documents through the profiles, we thus hope to be able to increase the number of documents judged by more than one user. The benefits of various methods will be studied here.

Résumé

Le but de cet article est de discuter de l'intégration de différents algorithmes d'analyse de liens au sein du projet GALILEI. En effet, dans le cadre de ce projet, visant à regrouper des utilisateurs en communautés virtuelles en fonction de leurs centres d'intérêts, on remarque que le système converge plus rapidement lorsqu'un document est consulté par plus d'un utilisateur. Sachant que les méthodes d'analyse de liens pourraient faire ressortir des documents « références » à travers les profils, on espère ainsi augmenter le nombre de documents jugés par plus d'un utilisateur. Pour cela, nous allons utiliser plusieurs méthodes.

Mots-clés : hyperliens, liens, graphe, hub, authority.

1. Introduction

Dans cet article, nous présentons différents moyens pour améliorer le groupement de profils en utilisant l'analyse des structures hyperliées. Tous les travaux ont été effectués dans le cadre du projet GALILEI[1], décrit dans la section 2. La section suivante évoque la notion d'hyperliens et les extensions qu'on peut y apporter. Dans la section 4, les concepts des différents algorithmes d'analyse d'hyperliens sont présentés et brièvement expliqués. L'intégration de ces méthodes dans le projet est abordée dans la section 5 et la méthodologie de validation utilisée dans GALILEI est ensuite présentée (section 6). La section 7 décrit les expériences qui ont été réalisées sur les différentes méthodes de calcul de liens ; puis elle compare les résultats obtenus. L'article se termine par un point sur les développements futurs (section 8) et une conclusion (section 9).

2. Le projet GALILEI

La quantité d'information, disponible électroniquement, a considérablement augmenté, vu le nombre sans cesse croissant de systèmes orientés documents, tel qu'Internet (Lawrence et Giles, 1998 et 1999). Afin de pouvoir traiter cette masse d'informations, il est de plus en plus

[1] Ce projet est subventionné par la Région wallonne, sous le contrat 01/1/4675, et est disponible à l'adresse suivante : http ://galilei.ulb.ac.be

nécessaire de s'armer d'un système de recherche documentaire, qui facilite le processus d'extraction de l'information en la structurant et en la prétraitant. Un article récent a rappelé que trouver l'information pertinente est devenu un problème crucial (Fogarty et Bahls, 2002).

Le but principal du projet GALILEI (Francq, 2003) est de cerner, le plus précisément possible, les centres d'intérêts des utilisateurs afin de pouvoir ensuite les regrouper correctement. Les utilisateurs doivent tout d'abord se créer des profils, un par centre d'intérêt. Ils peuvent ensuite commencer à émettre des avis sur la pertinence[2] des documents qu'ils consultent. Ces derniers sont ajoutés au profil, de même que le jugement correspondant.

Le système calcule la description des profils en se basant sur le contenu des documents, ainsi que sur les jugements émis par les profils.

Le système regroupe ensuite les différents profils en fonction de leur description : les profils ayant une description similaire sont regroupés afin de constituer un certain nombre de **communautés virtuelles**. Ce processus est accompli de façon dynamique par un algorithme génétique spécialement développé à cet effet, le GVCA « *Genetic Virtual Community Algorithm* ». Celui-ci se base sur la similarité entre profils, mais aussi et surtout sur deux types de contraintes dures, à savoir le taux de comportement en accord et en désaccord :

Le taux de comportement en accord représente le rapport entre le nombre de documents jugés comme pertinents par deux profils et le nombre total de documents jugés par ces deux profils. Pour chaque paire de profils, p_k et p_l, le taux de comportement en accord, $0 <= B_{same} <= 1$, est défini par : $B_{same}(p_k, p_l) = \frac{|D^I_{k,l}|}{|D_{k,l}|}$ où $D_{k,l}$ représente le sous-ensemble de documents jugés par les deux profils p_k et p_l et où $D^I_{k,l}$ représente le sous-ensemble de documents jugés de la même façon par les deux profils.

Le taux de comportement en désaccord représente le rapport entre le nombre de documents jugés différemment par deux profils et le nombre total de documents jugés par ces deux profils. Pour chaque paire de profils, p_k et p_l, le taux de comportement en désaccord, $0 <= B_{diff} <= 1$, est défini par : $B_{diff}(s_k, s_l) = \frac{|D^D_{k,l}|}{|D_{k,l}|}$, où $D^D_{k,l}$ représente le sous-ensemble de documents jugés différemment par les deux profils.

Lorsque le taux de comportement en accord est supérieur à un seuil fixé, on force les deux profils à être groupés ; de même, lorsque le taux de comportement en désaccord est supérieur à un certain seuil, on force les deux profils à ne pas être groupés dans la même communauté virtuelle. On voit donc que ces deux contraintes dures jouent un rôle important dans le groupement, mais pour que celles-ci aient un sens, il faut que les profils aient jugé un nombre minimum de documents (par exemple dix).

Une fois ces communautés virtuelles définies, l'information pertinente peut être diffusée entre les différents membres de ce groupe. Par exemple, certains documents considérés comme pertinents par un grand nombre d'utilisateurs d'une même communauté virtuelle peuvent être mis en partage pour toute la communauté.

Le système a été validé sur plusieurs collections de documents grâce à une méthodologie bien définie ; les résultats obtenus sont très prometteurs (Francq, 2003). Une solution informatique, disponible sous licence GNU GPL, a été développée.

[2] La pertinence d'un document peut être jugée de trois façons différentes : « *Relevant* » : le document est pertinent, « *Fuzzy Relevant* », le document n'est pas pertinent mais appartient quand même au centre d'intérêt et « *Irrelevant* » : le document n'appartient pas au centre d'intérêt ; et ce juste par un click de souris sur le document.

3. Extension de la notion d'hyperliens

Les hyperliens dans les documents HTML sont bien connus : ils permettent à l'utilisateur de naviguer d'une ressource à une autre, juste en cliquant sur ceux-ci. Cette notion d'hyperliens peut être étendue à une multitude d'autres types de documents ne contenant pas explicitement d'hyperlien. En effet, ceux-ci doivent juste contenir une structure qui apparaît dans un grand nombre de ces documents et qui permet de relier au moins deux ressources. Par exemple, les cocitations — bibliographie — des articles PDF peuvent être vues comme des liens entre deux ressources distinctes. Ces types de liens ne permettent pas de retrouver directement l'emplacement du document, mais sont tout de même capables de créer des liens entre différentes ressources. Dans la suite de l'article, tous les raisonnements sont fondés sur les principes des hyperliens ; ils peuvent bien sur être généralisés à n'importe quel autre type de liens.

L'utilisation des liens a pour but est d'améliorer et d'élargir plus rapidement une collection de documents. En effet, les liens rencontrés dans les documents contiennent une certaine quantité d'information intéressante, appartenant généralement au même centre d'intérêt, puisque l'auteur prend la peine de citer les autres ressources.

4. Aperçu des algorithmes de liens

Soit un ensemble de documents Web se citant mutuellement, cette collection peut être vue comme un graphe dirigé. Les algorithmes d'analyse de liens construisent la matrice d'adjacence, reflet du graphe basé sur le modèle de citation utilisé. Par exemple, le lien qui lie deux documents i et j sera représenté par la valeur 1 de l'élément W_{ij}. Les pages les plus intéressantes peuvent ensuite être extraites en calculant les vecteurs propres du système. Au sens de Kleinberg (1998), ces pages peuvent être divisées en deux catégories :

les hubs : pages contenant peu d'informations pertinentes, mais beaucoup d'hyperliens.

les authorities : pages contenant peu de liens, mais beaucoup d'informations pertinentes.

4.1. *L'algorithme HITS*

Kleinberg (1998) a proposé un algorithme, appelé HITS, capable d'identifier les meilleurs *hubs* et *authorities* au sein d'une collection hyperliée. Cet algorithme exploite la structure du graphe du web. Chaque document est vu comme un nœud du graphe dirigé et tout lien entre deux documents est interprété comme une arête entre deux nœuds. Partant d'une requête spécifique, appelée σ, l'algorithme crée tout d'abord un sous-graphe, S_σ ; il calcule ensuite le poids des *hubs* et *authorities* pour chaque nœud de S_σ.

Le principe utilisé par l'algorithme HITS est le suivant : un document a un poids *authority* élevé s'il est pointé par beaucoup de documents ayant un poids *hub* élevé et, vice versa, un document a un poids *hub* élevé s'il pointe vers beaucoup de documents ayant un poids *authority* élevé.

Plus précisément, en partant d'une série de documents hyperliés, l'algorithme HITS construit le graphe dirigé, associé à la collection. Idéalement, la collection S doit respecter les propriétés suivantes :

(i) S est relativement petit.
(ii) S contient beaucoup de pages pertinentes.
(iii) S contient la plupart des meilleurs *authorities*.

Le graphe est représenté par une matrice d'adjacence, W, de taille $n \times n$, où n représente le nombre de documents utilisés. L'élément W_{ij} prend la valeur 1 s'il existe une arête entre les

nœuds i et j dans le graphe dirigé, et 0 dans les autres cas. Généralement, la troisième condition n'est pas satisfaite et la collection S doit être étendue en explorant un certain nombre de liens du graphe (Kleinberg, 1998). L'algorithme peut ensuite calculer les relations de renforcement mutuel entre les *hubs* et les *authorities* en itérant les règles de mise à jour suivantes :

$$h_i^{(k+1)} = \sum_{j:(j \to i)}^{n} a_j^{(k)} \ ; \ a_i^{(k+1)} = \sum_{j:(i \to j)}^{n} h_j^{(k)} \tag{1}$$

où « $i \to j$ » signifie que le document i pointe vers le document j. Les équations ci-dessus peuvent aussi s'écrire en utilisant la notation matricielle :

$$\begin{cases} h^{(k+1)} & = & Wa^{(k)} & = & (WW^T)h^{(k)} \\ a^{(k+1)} & = & W^T h^{(k)} & = & (W^T W)a^{(k)} \end{cases} \tag{2}$$

En initialisant les vecteurs poids avec un vecteur colonne unité $n \times 1$: $h^{(0)} = a^{(0)} = [1, 1, 1, ...]^T$ et en normalisant les vecteurs résultant de l'application des règles de mise à jour, par une norme euclidienne, nous utilisons la « *power method* »[3], qui converge vers le vecteur propre principal de la matrice symétrique. Kleinberg a prouvé que h converge vers le vecteur propre principal normalisé de la matrice WW^T, alors que a converge vers le vecteur propre normalisé de la matrice $W^T W$.

4.2. *L'algorithme SALSA*

Dans cette section, nous introduisons SALSA, « *the Stochastic Approach for Link-Structure Analysis* », un algorithme basé sur la théorie des chaînes de Markov proposé par Lempel et Morgan (2000 et 2001). L'algorithme utilise les propriétés d'une marche aléatoire effectuée sur une collection de documents hyperliés. Tout comme l'algorithme de Kleinberg, SALSA commence par construire une collection de base « *base set* » issue graphe des liens. SALSA est basé sur l'intuition qu'une page « *authoritative* » doit être visible de beaucoup de pages du set de données. Une marche aléatoire dans ce graphe visitera donc, avec une probabilité relativement élevée, un certain nombre d'*authorities*. La théorie des marches aléatoires est combinée à la notion de *hubs* et *authorities*, ce qui nous mène à la nécessité d'analyser deux chaînes de Markov différentes : la chaîne des *hubs* visités et la chaîne des *authorities* visitées, ce qui nous donne, pour chaque page, deux poids distincts : celui des *hubs* et celui des *authorities*. Pour générer les états de transition de chacune de ces chaînes, deux arêtes du graphe doivent être traversées, la première vers l'avant (en suivant un lien sortant), la seconde vers l'arrière (en suivant un lien rentrant) ou vice versa. Les poids des *authorities* sont définis comme étant la distribution stationnaire de la chaîne explorant en premier un lien aléatoire vers l'arrière et ensuite un vers l'avant, alors que les poids des *hubs* sont définis comme étant la distribution stationnaire de la chaîne explorant en premier un lien aléatoire vers l'avant et ensuite vers l'arrière.

Plus précisément, partant d'une collection de documents hyperliés, nous pouvons construire le graphe dirigé G. Soit $Back(i) = \{k : k \to i\}$, l'ensemble des nœuds qui pointent vers i, par

[3] La « *power method* » est une méthode numérique itérative pour le calcul du vecteur propre dominant d'une matrice symétrique pour le problème aux valeurs propres suivant :

$$\begin{cases} h & \propto & Wa \\ a & \propto & W^T h \end{cases}$$

exemple les nœuds qui peuvent être atteints à partir de i en suivant un lien en arrière, et soit $Forw(i) = \{k : i \rightarrow k\}$, l'ensemble de tous les nœuds qui peuvent être atteints en partant de i en suivant un lien vers l'avant. $|Back(i)|$ représente le nombre de nœuds qui pointent vers i, tout comme $|Forw(i)|$ représente le nombre de nœuds vers lesquels i pointe.

Nous pouvons maintenant définir les deux matrices stochastiques, qui contiennent les probabilités de transition pour les chaînes de Markov respectivement pour les *hubs* et les *authorities*.

1. La matrice pour les *hubs*, H :

$$h_{i,j} = \sum_{k:k \in Forw(i) \cap Forw(j)} \frac{1}{|Forw(i)|} \frac{1}{|Back(k)|} \tag{3}$$

2. La matrice pour les *authorities*, A :

$$a_{i,j} = \sum_{k:k \in Back(i) \cap Back(j)} \frac{1}{|Back(i)|} \frac{1}{|Forw(k)|} \tag{4}$$

Un élément $a_{i,j} > 0$ signifie qu'au moins un nœud k pointe vers les deux nœuds i et j. Le nœud j est donc atteignable en partant du nœud i en deux étapes : la première en remontant le lien $i \rightarrow k$, la deuxième en suivant le lien $k \rightarrow j$.

Les équations (3) et (4) peuvent être écrites sous forme matricielle. On obtient dès lors un système similaire à celui obtenu en (2), pouvant aussi être résolu grâce à la « *power method* ».

5. Intégration des méthodes de liens dans GALILEI

Les utilisateurs jugent chacun un ensemble de documents. Tous les profils du système sont donc composés d'ensembles de documents. Le système va ensuite appliquer les algorithmes d'analyse de liens pour chaque ensemble de documents pertinents en étendant les graphes, comme dans la section 4.1. Le but de ces algorithmes est d'affecter à chaque document un poids *hub* et un poids *authority*. Les meilleures pages de chacune des catégories sont ajoutées aux profils. On espère ainsi que les documents les plus intéressants se retrouveront dans les profils correspondant au centre d'intérêt. De plus, les documents de type *hub*, pointent vers des documents contenus dans les profils ; on espère ainsi que deux profils, n'ayant pas jugé de documents communs mais appartenant à une même communauté virtuelle, se verront attribuer un même document *hub*. Il sera dès lors encore plus facile de les regrouper.

6. Méthodologie de validation de GALILEI

La méthodologie de validation définit des cycles de tests simulant le comportement des utilisateurs (Francq, 2003).

Seuls deux types de tests, permettant de quantifier l'apport des méthodes de liens, seront exposés ici. Il s'agit, d'une part, d'un test de simulation du système réel et, d'autre part, de tests permettant de mesurer les différents taux de comportement, exposés précédemment.

6.1. *Set de données et catégorisation*

Les différents tests seront effectués sur une collection de documents précatégorisés par centre d'intérêt. Cette collection a été construite par l'aspiration de plusieurs sites traitant de différents sujets. L'ensemble du set de données contient 5935 documents et est composé de six catégories, comme le montre le tableau 1.

6.2. *Simulation du système réel*

L'objectif est de simuler l'évolution du système en fonctionnement réel pendant une période de 200 unités de temps. Les profils doivent être calculés et regroupés en communautés virtuelles afin de pouvoir mesurer la convergence et la stabilité du système. On part du principe que chacune des catégories correspond à une **communauté virtuelle idéale**. Le but recherché est que le système simulé puisse converger vers le groupement idéal, c'est à dire que chaque communauté virtuelle corresponde à un centre d'intérêt prédéfini. Le test commence par l'initialisation et répète ensuite aléatoirement, jusqu'à la fin de la période, une phase de création ou de rétroaction.

Initialisation : deux profils issus de deux catégories différentes sont créés. On simule ensuite une série de jugements sur des documents (en respectant un certain pourcentage d'erreurs par rapport à la situation idéale). Le système calcule ensuite les deux profils et les communautés virtuelles. Le résultat est ensuite comparé au groupement idéal.

Rétroaction : pendant une ou plusieurs unités de temps, certains documents pertinents sont partagés entre les utilisateurs d'une même communauté. Le système recalcule ensuite les profils et les communautés virtuelles. Le résultat est ensuite comparé au groupement idéal.

Création : on crée un nouveau profil associé à une catégorie aléatoire parmi celles prédéfinies. Ceci permet de simuler l'arrivée d'un nouvel utilisateur dans le système et éventuellement l'arrivée d'un nouveau centre d'intérêt si celui-ci appartient à une nouvelle catégorie. On génère pour ce profil un ensemble de jugements sur des documents. Les profils et communautés virtuelles sont recalculés. On compare ensuite le groupement calculé au groupement idéal.

6.3. *Mesures*

***adjusted RAND Index* :** A chaque unité de temps, l'*adjusted RAND Index* (Hubert et Arabie, 1985) du système simulé est calculé. Cette valeur, comprise entre -1 et 1, représente la qualité de la réponse par rapport au groupement idéal. Lorsque cette valeur est nulle, le groupement calculé est aléatoire. Ce test permet d'observer l'évolution du système, ainsi que d'en déduire sa convergence et sa stabilité.

Nous avons vu précédemment (section 2), que d'autres critères permettent de quantifier la solution de l'algorithme de groupement. Afin de pouvoir relever ces différentes mesures, nous prenons une photographie du système à un instant t donné. Ce processus est réalisé une dizaine de fois de manière à avoir un aperçu du comportement du système à différents moments.

Nombre moyen de profils ayant jugé un même document : Ce test simule le système et

N°	Centre d'intérêt	Nbr
1	C++	1283
2	Html	535
3	Java	1741
4	Linux	972
5	Perl	368
6	Windows	1046

Tableau 1. Description du set de données

prend une configuration particulière au temps t ; il calcule, pour chaque document connu du système, le nombre de profils ayant émis un jugement sur le document en question. La moyenne des valeurs obtenues est ensuite calculée. Ce test n'a de sens que s'il est exécuté par différentes méthodes, afin que les courbes puissent être comparées entre elles. Il donne un aperçu du recouvrement entre profils et donc une bonne estimation de la facilité à les grouper. La comparaison des résultats sera discutée dans la section 7.2.

Taux moyen de comportement en accord : Ce test se base sur le taux de comportement en accord. Comme vu précédemment dans la section 1.3, plus la valeur de ce paramètre est élevée, plus le groupement sera aisé, puisque plus de profils auront jugé les même pages. Ce test simule le système, prend une configuration à un temps t et calcule la moyenne sur les différents profils. La comparaison des résultats sera discutée dans la section 7.3.

Taux moyen de comportement en désaccord : Ce test se base sur le taux de comportement en désaccord. Tout comme le taux de comportement en accord, plus la valeur de ce paramètre est élevée, plus le groupement est facilité, puisque ces deux profils ne devront pas être groupés ensemble. Le test se déroule de la même façon que le test exposé ci-dessus. Les résultats sont commentés dans la section 7.4.

6.4. *Résultats antérieurs*

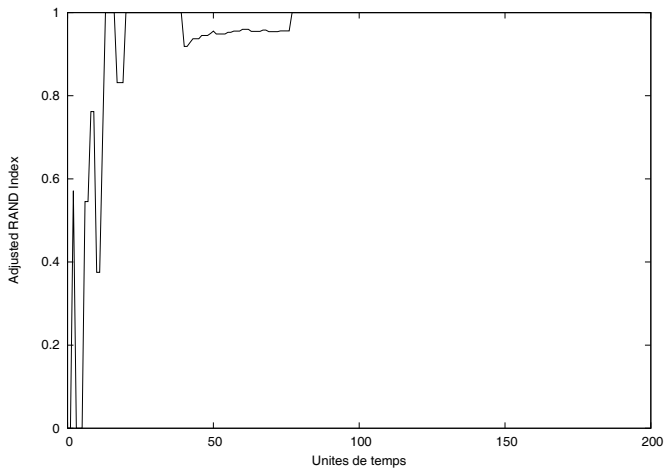


Figure 1. Résultat de la simulation du système réel

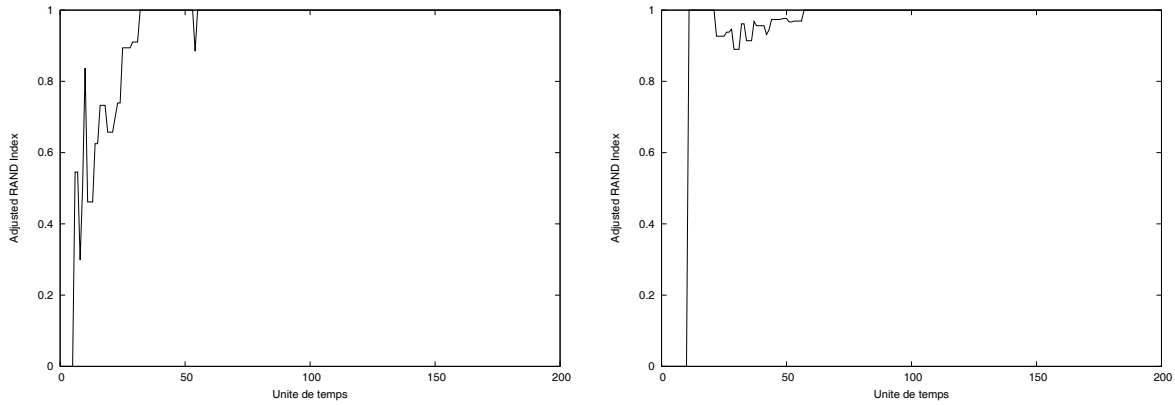
La figure 1 représente la simulation du système réel sans l'analyse des liens. On remarque clairement que le système est fortement instable lors des premières unités de temps. En effet, le peu de documents jugés par les utilisateurs ne donne pas suffisamment d'informations pour permettre aux algorithmes de grouper correctement les différents profils lorsque de nouveaux profils sont créés. Le système va ensuite converger vers le groupement idéal, mais il lui faudra encore quelques unités de temps avant de devenir complètement stable.

7. Méthode de validation pour les algorithmes de liens

Les différentes expériences ont pour but de quantifier l'apport des méthodes de liens au système. Les tests seront toujours effectués avec et sans les méthodes d'analyse de liens, afin de pouvoir comparer les résultats.

7.1. *Simulation réelle du système*

La figure 2(a) montre les résultats de la simulation du même système qu'à la section 6.2, mais utilisant en plus l'algorithme de liens HITS. En comparant la figure 2(a) à la figure 1, on peut tirer un certain nombre de conclusions. Le système présente de moins bons résultats lors des premières itérations, ce qui est dû au non respect des conditions initiales de la méthode (voir section 4.1.). En effet, les profils ne contiennent pas assez de documents pour construire un graphe dirigé correct et les documents ainsi rajoutés perturbent le groupement. Par la suite, le système converge de la même manière, mais il devient stable plus rapidement.



(a) avec la méthode HITS (b) avec la méthode SALSA

Figure 2. Simulation du système réel

En comparant la figure 2(b), qui présente les résultats de la simulation en utilisant SALSA, à la figure 1, on constate que lors des dix premières itérations la méthode fausse complètement le groupement, pour la même raison que celle évoquée précédemment. Passé ce cap, le groupement converge beaucoup plus rapidement et de façon beaucoup plus stable.

En comparant les deux figures 2(a) et 2(b), on constate que, sur cette collection de documents, la méthode SALSA donne de meilleurs résultats, puisque le système converge plus vite et qu'il reste plus stable.

7.2. *Nombre moyen de profils ayant jugé un même document*

Plus les valeurs de ce test sont importantes, plus les groupes seront déterminés facilement. Les deux courbes ainsi obtenues (voir figure 3) montrent l'impact des méthodes de liens sur le système. En effet, grâce à l'analyse des liens, d'autres documents considérés comme pertinents sont rajoutés aux profils. On constate donc un accroissement du nombre de pages jugées en commun par plusieurs profils, et donc un groupement plus aisé.

7.3. *Taux moyen de comportement en accord*

On a vu précédemment (section 2) que si le taux de comportement en accord entre deux profils est supérieur à un seuil imposé, alors ces deux profils issus d'un même centre d'intérêt sont mis dans la même communauté virtuelle. Comme le montre la figure 3, le taux moyen de comportement en accord augmente quelle que soit la méthode de liens utilisée. Ces résultats indiquent bien que trouver le meilleur groupement sera plus aisé lorsque l'on adjoint au système les méthodes d'analyse de liens.

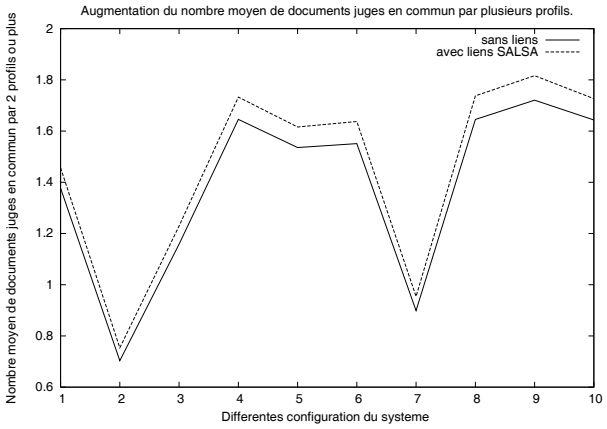


Figure 3. Average Ratio

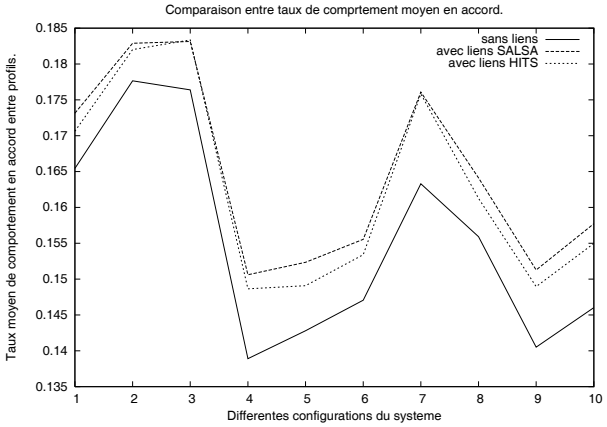


Figure 4. Taux de comportement en accord

7.4. *Taux moyen de comportement en désaccord*

Comme vu précédemment (section 2), si le taux de comportement de désaccord entre deux profils est supérieur à un seuil imposé, alors les deux profils, issus de centres d'intérêt différents, ne seront pas placés dans la même communauté virtuelle.

On voit clairement, sur la figure 4, l'amélioration obteneu lorsque l'on utilise les méthodes d'analyse de liens. Ceci s'explique facilement par l'ajout automatique de documents aux profils. En effet, deux profils venant de communautés virtuelles différentes recevront chacun des documents propres à leur centre d'intérêt. Cet ajout d'informations séparera d'autant plus les deux profils et le groupement en sera d'autant plus aisé.

8. Développements futurs : l'algorithme hybride

Nous remarque que les deux méthodes d'analyse de liens donnent plus ou moins le même genre de résultats quel que soit le type de test, pour la collection de documents utilisée pour les tests. D'autres collections de documents ne donnent pas des résultats aussi homogènes pour les différentes méthodes. C'est pourquoi une méthode hybride est en cours de développement.

Les deux algorithmes, présentés à la section 4, ont certes une série de propriétés très intéressantes, mais ont aussi leurs propres limites. Un des problèmes les plus indésirables rencontrés avec l'algorithme HITS est ce qu'on appelle le « *topic drift* » : l'algorithme converge vers la communauté la plus hyperliée au sein du set de base, mais cette communauté n'a parfois

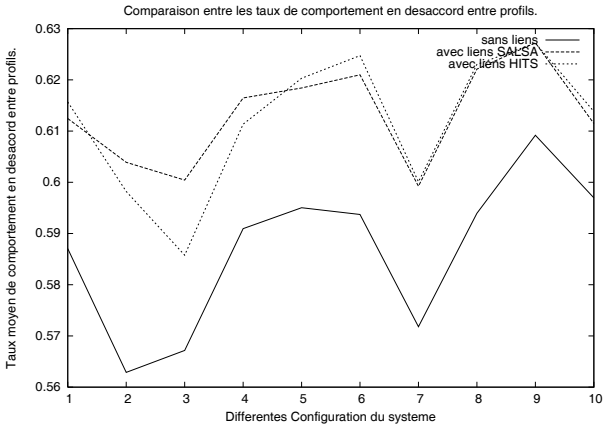


Figure 5. Taux de comportement en désaccord

rien à voir avec le centre d'intérêt du groupe. SALSA peut être vu comme un premier moyen de contourner ce type d'instabilité ; certaines modifications, détaillées ci-dessous, peuvent encore améliorer les résultats.

L'idée retenue consiste à appliquer un seuil durant la phase de calcul des *hubs* et *authorities*, afin de ne tenir compte que des pages dont le poids est élevé ; cela revient à dire que lors de la phase de calcul on ne garde que les documents considérés comme bon *hub* ou bonne *authority*. Nous détaillerons ici brièvement la méthode que l'on a appelé « *Authority-Mean-Treshold*», laquelle consiste à appliquer un seuil lors du calcul des poids des *hubs*. Lors de ce calcul, pour une page i, l'algorithme ne prend pas en compte toutes les pages pointées par i mais seulement celles qui font partie des meilleures (le nombre de ces meilleurs pages est représenté par le seuil).

Cette philosophie peut aussi bien être appliquée lors de la phase de calcul des *authorities*, ainsi que lors du calcul des *hubs* et des *authorities* simultanément. Cet algorithme sera paramétrable de manière à pouvoir choisir la valeur des seuils, ainsi que le moment où de leur application.

9. Conclusions

Nous avons intégré plusieurs algorithmes d'analyse de liens au sein du projet GALILEI, le but premier étant d'utiliser ces liens pour déterminer les meilleurs documents *hub* et *authority* d'un profil. Ces documents sont ensuite proposés au profil en question dans l'espoir qu'ils soient du même centre d'intérêt. Les résultats nous confirment la validité des algorithmes. Le nombre de documents associés aux profils augmente d'avantage et les documents ainsi ajoutés sont pertinents par rapport aux centres d'intérêt du profil. Les communautés virtuelles sont donc plus vite définies et de façon plus précise. Le système convergera et se stabilisera plus rapidement.

Références

Fogarty M. et Bahls (2002). Information Overload : Feel the pressure ? *The Scientist*, vol. (16/16).

Francq P. (2003). *Structured and Collaborative Search : An integrated approach to share documents among users*. PhD Thesis. Université Libre de Bruxelles.

Hubert L. et Arabie P. (1985). Comparing partitions. *Journal of Classification* : 193-218.

Kleinberg J.M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithm* : 668-677.

Lawrence S. et Giles C.L. (1998). Searching the World Wide Web. *Science*, vol. (280) : 98-100.

Lawrence S. et Giles C.L. (1999). Accessibility of information on the Web. *Nature*, vol. (400) : 107-109.

Lempel R. et Morgan S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In *9th International World Wide Web Conference*.

Lempel R. et Morgan S. (2001). The Stochastic approach for link-structure analysis. *Transaction on Information Systems*, vol. (19/2) : 131-160.

El uso de la estadística en la didáctica de las lenguas extranjeras con fines específicos : descripción del proceso de selección del léxico típico del discurso económico empresarial en español

Lieve Vangehuchten

Universiteit Antwerpen – Prinsstraat 13 – 2000 Antwerpen – België
lieve.vangehuchten@ua.ac.be

Abstract

This research was triggered by my observation that the lexical content of currently existing didactic material for the purpose of the acquisition and teaching of Economic and Business Spanish is composed on the basis of merely intuitive criteria and, what is more, that it is the work of laymen, without consulting specialists (economists). This paper builds on the outcome of my PHD-research in which I designed a method to enable the objective selection of vocabulary. The method itself is based on the insights of modern terminology with regard to the identification of the terms as well as on some lexico-statistical techniques, and was tested on a sample-corpus of Applied Economics Spanish, counting 118.365 occurrences. Given the modest dimension of this corpus the results are rather indicative than representative.

Keywords : corpus linguistics, statistics for terminological purposes, terminology, lexicology, second language acquisition research.

Resumen

El punto de partida de este trabajo es la constatación de que el contenido léxico de las herramientas de aprendizaje del español económico empresarial es, en la gran mayoría de los casos, el resultado de una selección subjetiva e intuitiva. En el marco de la elaboración de una tesis doctoral, se desarrolló una metodología para la identificación y la selección científica y objetiva del léxico típico de una lengua especializada. En cuanto a la identificación, se siguió la teoría de la corriente lingüística de la Terminología que estimula la investigación semasiológica en corpus. Para el proceso de selección se ha recurrido a unas técnicas léxico-estadísticas que se han aplicado a un corpus modélico de 118.365 ocurrencias. Dado el tamaño modesto de este corpus de discurso económico empresarial, los resultados son más bien indicativos que representativos.

Palabras clave : lingüística de corpus, léxico-estadística, terminología, lexicología, lingüística aplicada a la adquisición de una lengua extranjera.

1. Introducción

El punto de partida de este trabajo es la constatación de que el contenido léxico de las herramientas de aprendizaje del español económico empresarial es, en la gran mayoría de los casos, el resultado de una selección subjetiva e intuitiva (Vangehuchten, 2003). Con vistas a la elaboración de una metodología para la selección científica y objetiva del léxico típico de una lengua especializada, se ha recurrido a algunas técnicas estadísticas que han sido aplicadas a un corpus modélico de discurso económico empresarial. En este artículo se comentarán primero la composición del corpus modélico así como su norma de lematización. Después se describirán las distintas categorías léxicas presentes en el corpus y la manera de la que la estadística contribuye a su clasificación y selección didáctica.

2. Presentación del corpus modélico

Se trata de un corpus de discurso didáctico, o sea un texto medianamente especializado, de comunicación entre un especialista y aprendices. Es un manual de introducción a la economía empresarial de Eduardo Pérez Gorostegui, catedrático de economía de la empresa Universidad Nacional de Educación a Distancia (España). Se podría reprochar a este corpus modélico el hecho de ser de un solo autor, por lo cual no sería representativo del discurso económico empresarial en general, sino del idiolecto del autor. De ahí que resulte necesario poner en claro que el corpus modélico no sirve como base para la confección de un glosario o diccionario de léxico económico empresarial, sino que desempeña la función de objeto de experimentación. Por otra parte, cabe señalar que el texto pertenece a un registro formal y escrito de lenguaje técnico sometido a un grado elevado de normalización. Asimismo, se trata de un texto de instrucción, un manual que tiene un uso generalizado en las universidades españolas. Todas estas características nos permiten decir que el carácter idioléctico de este texto está muy limitado y se manifiesta sin duda más en algunas estructuras sintácticas recurrentes del autor, que en su empleo del léxico técnico.

La norma de lematización ha sido la unidad léxica (UL), por lo cual el resultado es más fino que el de un *type*/*token* análisis. Una UL es la realización de un lexema, que se puede componer de más de una sola forma y que puede sufrir variaciones gramaticales (sintácticas y morfológicas) con tal de que no afecten su significado básico. El significado de una UL es, por tanto, único e invariable, aunque permite empleos pragmáticos individuales. En total se han identificado 4.933 UL distintas sin contar los nombres propios. La lematización se ha hecho tanto automática como manualmente. Para la parte automática de la lematización del corpus, hemos recurrido a un lematizador español (*RELAX* Part-of- Speech *tagger*[1]). Ha sido imprescindible corregir los resultados manualmente, dado que este *tagger* fue desarrollado para lematizar la lengua general y no es capaz de distinguir entre significados especializados y generales, ni tampoco de reconocer términos compuestos.

El tamaño del corpus analizado sólo es de 118.365 ocurrencias (OC). No obstante, según *The Handbook of Terminology Management* un corpus de 100.000 formas léxicas ya es suficiente para sacar conclusiones significativas, dado que el léxico utilizado para tratar una temática especializada es más restringido que el léxico de un discurso no especializado :

> "*As a rule of thumb, special-language corpora already start to become useful for key terms of the domain in the tens of thousands of words, rather than the millions of words required for general-language lexicography*." (Ahmad y Rogers, 2001 : 736)[2]

Asimismo, la representatividad del corpus se puede demostrar estadísticamente al comparar la cantidad de UL nuevas por capítulo o el crecimiento del léxico a través del corpus, como se desprende del gráfico 1. Se puede ver que la evolución del léxico nuevo es bastante lógica : la tasa de UL nuevas disminuye progresivamente entre los capítulos 1 y 7, sufre altibajos entre los capítulos 7 y 10 para llegar a una situación de estabilidad relativa a partir del capítulo 11. Clijsters (1990 : 84) llama a este punto 'la saturación' del léxico :

[1] Este *tagger* fue desarrollado por el *Natural Language Processing Group* de la Universitat Politècnica de Catalunya, en colaboración con el *Computational Linguistics Laboratory* de la Universidad de Barcelona (información en http ://www.lsi.upc.es/~nlp/).

[2] En este contexto cabe citar también el estudio de Sutarsyah *et al*., (1994), quienes, para la comparación del vocabulario de la lengua económica en general – o sea, no sólo empresarial – con el vocabulario de la lengua académica, se sirven como corpus de un manual de economía general de un solo autor que cuenta 300.000 *tokens*.

> " […] *l'accroissement en vocables nouveaux tend vers un point de saturation (V n'augmente plus qu'à grand-peine)* […] "

Este punto indica que se ha empezado a agotar el léxico del asunto tratado y que habrá un momento en que el inventario de UL nuevas tocará fondo. Aún según Clijsters, este momento de inicio de agotamiento representado por el punto de saturación también es una prueba de la representatividad del corpus. Esto significa, como se desprende claramente del gráfico, que un corpus compuesto de los siete primeros capítulos del manual – 51.028 OC en total – no hubiera sido representativo, dado que después del séptimo capítulo el punto de estancamiento todavía no se alcanza. El punto de saturación parece coincidir con el capítulo 11, o sea después de 79.363 OC. Con una totalidad de 118.365 OC, el corpus supera por tanto con creces el punto de saturación.

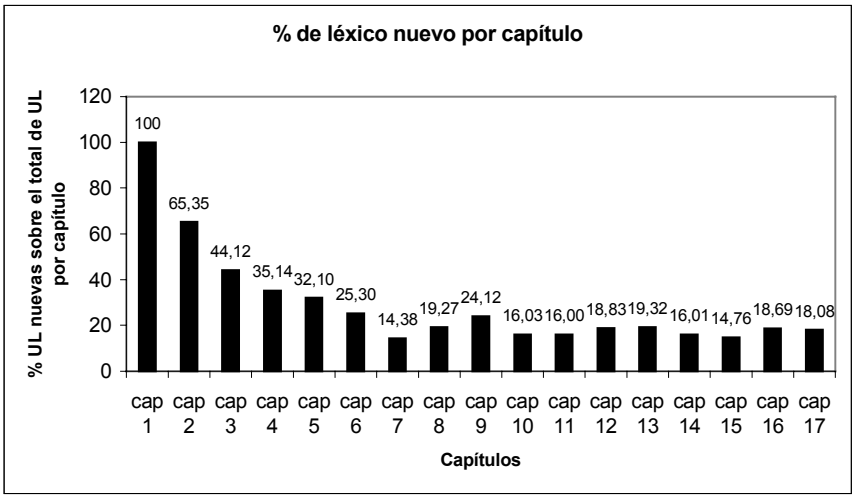


Gráfico 1.

3. Las categorías léxicas

En el siguiente cuadro se presentan las categorías léxicas del corpus :

	UL	**OC**
Funcionales (F)	5,07%	55,99%
Términos (T)	28,54%	12,22%
Términos auxiliares (TA)	5,37%	1,80%
Generales (G)	61,02%	29,99%
Total	**100%**	**100%**

Cuadro 1.

3.1. *El léxico funcional*

El léxico funcional manifiesta la tasa de cobertura más importante : el 55,99% de las OC. Sin embargo, sólo constituyen el 5,07% de todas las UL, o sea que la diversidad léxica en esta categoría es muy baja. El léxico funcional, que se llama también léxico gramatical o léxico vacío por su valor semántico casi inexistente, representa una clase finita de palabras muy

frecuentes. En cuanto a la selección didáctica del léxico funcional, queda claro que este grupo finito de palabras muy frecuentes debe formar parte de manera exhaustiva del contenido léxico de un curso de español como lengua extranjera.

3.2. *Los términos y los términos auxiliares*

3.2.1. *La identificación*

La identificación se ha hecho manualmente a partir de los criterios formulados por la corriente lingüística en la Terminología (Kocourek, 1982 ; Sager, 1990 ; Cabré, 1999 ; Temmerman, 2000), ya que todavía no existe un programa de extracción terminológica que produzca un resultado lo suficientemente preciso.

3.2.2. *La selección*

Hoy en día vuelve a admitirse que la frecuencia es un criterio eficaz y fiable para seleccionar y graduar el léxico en un contexto didáctico. Para dar resultados verídicos, el criterio de la frecuencia necesita ser combinado con el de la dispersión. La estadística léxica elaboró distintas fórmulas para combinar frecuencia y dispersión, siendo, sin duda, una de las más conocidas el '*usage coefficient'* de Juilland y Chang-Rodríguez (1964 : XVIII). Según Müller (1977 : 74) es importante tomar en consideración si el corpus está dividido en partes iguales o fragmentos desiguales. En el último caso, que es el nuestro, el padre de la estadística léxica sugiere una frecuencia "corregida" : KF = $\left(\sum \sqrt{p_i f_i}\right)^2$

En esta fórmula, que está basada en la media geométrica y que fue desarrollada por Rosengren (1971 :119), p_i representa el tamaño relativo del fragmento o capítulo *i* en que figura la UL en cuestión, o sea la totalidad de OC en el fragmento *i* dividida por la totalidad de OC en el corpus. La frecuencia de la UL en el fragmento *i* corresponde a f_i. Al calcular el producto de f_i con p_i para cada capítulo en que está el lema, se obtiene una cifra que Rosengren llama la frecuencia corregida, ya que esta fórmula tiene en cuenta la dispersión de cada lema para corregir su frecuencia. Por lo tanto, los lemas que ocurren en total dos veces en dos capítulos, se estiman más importantes que los lemas con una frecuencia superior pero concentradas en un solo capítulo.

Se ha demostrado que la fórmula de la frecuencia corregida de Rosengren ofrece un criterio fiable para ordenar los términos según su importancia. No obstante, conviene hacer dos observaciones respecto del ranking así obtenido. Una primera observación concierne la utilidad de someter el resultado exhaustivo al estudiante de español económico. Hazenberg (1994 : 47) demuestra, en su intento de establecer un vocabulario fundamental neerlandés para estudiantes extranjeros que desean empezar una carrera académica, que la frecuencia y la dispersión sólo son útiles en la medida en que la tasa de nuevas palabras a aprender sea compensada por una ganancia considerable en cuanto a la tasa de cobertura. Dicho de otro modo, a partir del momento en que el aprendizaje de nuevas palabras ya no corresponde a un progreso considerable en cuanto a la tasa de cobertura, no tiene sentido continuar imponiéndolas al estudiante, ya que se entra en la zona donde las palabras manifiestan todas la misma frecuencia bajísima, por lo cual la posibilidad de toparse con ellas en un texto representativo del lenguaje examinado queda muy reducida. Claro está que esto no significa que las UL que siguen a este punto no sean importantes, sino que vale más que a partir de este momento el estudiante centre sus intereses en un subcampo determinado y que extienda sus conocimientos léxicos con el vocabulario desconocido que éste contiene.

Si se aplica esta idea a los términos económicos y auxiliares del corpus, después de haberlos ordenado según su frecuencia corregida, es posible distinguir entre los términos básicos y los términos de continuación o subespecialización. A continuación se presentan las tasas de cobertura de los términos económicos y auxiliares en el corpus, teniendo en cuenta también los *hapax legomena* :

T	OC	% de cobertura	TA	OC	% de cobertura
0	0	0%	0	0	0%
200	10.803	74,68%	50	1.593	74,96%
400	12.419	85,85%	100	1.853	87,20%
600	13.213	91,34%	150	1.997	93,98%
800	13.727	94,89%	200	2.060	96,94%
1.000	14.058	97,18%	265	2.125	100,00%
1.200	14.258	98,56%			
1.408	14.466	100,00%			

Cuadro 2.

Los 200 primeros términos económicos más frecuentes y mejor repartidos, manifiestan una tasa de cobertura muy elevada del 74,68%. A partir de este punto la ganancia disminuye sin cesar : si de 200 a 400 UL todavía es de un 11,17%, baja a un 5,49% de 400 a 600 UL y a un 3,55% de 600 a 800 UL. Según Hazenberg (1994 : 48), al repartir las UL del corpus en partes equitativas por orden de frecuencia decreciente, una progresión del 3,5% aún se puede considerar sustancial en términos de comprensión potencial. Esto significa, para los términos económicos del corpus modélico, que 800 de ellos se pueden seleccionar como fundamentales en base a su frecuencia corregida. Después de este punto, el estudio de 200 UL más, sólo generaría una ganancia del 2,29%. Por consiguiente, hay 608 términos que quedan fuera de los términos seleccionados como fundamentales. Entre ellos figuran los 504 *hapax legomena*. Es necesario incluir 5 términos más por la siguiente razón. Dado que el término en la posición 800 en la lista por frecuencia corregida tiene el resultado 0.14, pero que hay 5 términos más con este mismo resultado que vienen después debido al orden alfabético, estos últimos tienen el mismo valor que el primero. Esto significa que se seleccionan, en total, 805 términos económicos. Por lo que se refiere a los términos auxiliares, la ganancia en cobertura disminuye considerablemente después de las 150 primeras formas – se reduce a un 2,96% entre las 150 y 200 UL, y a un 3,06% entre las 200 y 265 UL -, punto a partir del cual la selección en base a la frecuencia corregida deja de tener sentido.

3.3. *El léxico general*

La tasa de léxico general en el corpus es muy elevada (el 61,02% de las UL y el 29,99% de las OC). El objetivo de este apartado es examinar cómo se puede reconocer el léxico general típico del discurso económico empresarial académico. A este propósito se ha sometido el léxico general del corpus modélico a una comparación con el léxico de un corpus de español

general, el corpus Cumbre[3] (19.412.588 *tokens*). Concretamente se busca confirmación de las siguientes hipótesis :

1. la existencia de un léxico subeconómico, o sea UL con un significado orientado hacia el mundo de la economía, aunque no definidas como términos por los especialistas, por ejemplo *importe*, *suma*, *ascender*, etc.
2. la existencia de un léxico general básico, común al lenguaje general y al lenguaje económico. Este léxico general fundamental, se denomina '*core vocabulary*' en la literatura anglosajona (por ejemplo Dollerup *et al.,* 1989 : 22 ; Stubbs, 2001 : 42-43), que lo define como palabras muy frecuentes con una distribución semejante en una gran variedad de textos, estilísticamente neutras, semánticamente útiles por ser hiperónimos, e imprescindibles para comentar cualquier asunto.

A fin de comparar las frecuencias de UL de dos *corpora* de tamaño distinto, la estadística ha elaborado varias fórmulas. La más conocida es, sin duda, el *método del χ cuadrado o de Pearson*, pero éste requiere para el objetivo que nos proponemos frecuencias absolutas superiores a 5, lo que no es el caso de todas las UL generales en Econ. Una fórmula que no exige una frecuencia mínima es el denominado '*log-likelihood ratio*', aplicado a la comparación léxica de dos *corpora* por Rayson y Garside (2000). Este método parte de las frecuencias absolutas u observadas (O) de una palabra en dos corpus para calcular en primer lugar sus frecuencias teóricas o esperadas (E). Para ello se necesita conocer el número de formas léxicas en cada corpus (N) :

$$E_i = \frac{N_{ii} \sum_i O_i}{\sum_i N_i}$$

Las frecuencias esperadas de la palabra i corresponden al producto de N_i o el total de OC en el corpus concernido, con la suma de las frecuencias observadas dividido por la suma de la totalidad de OC en los dos corpus. Una vez que se conocen las frecuencias esperadas de la palabra, el '*log-likelihood ratio*' se calcula según la siguiente fórmula :

$$LL = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

El '*log-likelihood ratio*' (LL) se obtiene al multiplicar por dos la suma del producto de cada frecuencia observada con el logaritmo de la división de esta misma frecuencia observada por su frecuencia esperada. El valor obtenido expresa la diferencia entre la frecuencia en el corpus de base y la del corpus de referencia. Cuanto más alta la cifra, más significativa la diferencia. Los resultados obtenidos mediante la fórmula de '*log-likelihood ratio'* son estadísticamente significativos a partir de 3,84 cuando alpha = 0.05 y sólo a partir de 6.63 cuando alpha = 0.01 ($\chi^2\ \alpha[1]$ = 3,84/6,63 para un test unilateral, Kanji 1993 : 168).

Al aplicar la fórmula de *log-likelihood* a las frecuencias absolutas y esperadas de las UL generales se obtiene una cifra para las UL generales que figuran en ambos corpus. Dado que en lingüística se suele adoptar un umbral de probabilidad de error de 1 sobre 20 o del 5%,

[3] El corpus, compuesto por A. Sánchez *et al.* y cuyo léxico fue puesto a nuestra disposición gracias a su amabilidad, cuenta con 54.737 *types* que se realizan 19.412.588 veces y "es ampliamente representativo de la lengua española en España e Hispanoamérica (sin olvidar las áreas hispanohablantes de Estados Unidos), en su variedad escrita y oral, y en géneros y ámbitos variados" (2001 : 8).

incluimos entre las UL características del lenguaje económico empresarial académico todas las palabras cuya frecuencia relativa en Econ es superior a la que aparece en Cumbre y que tienen un resultado '*log-likelihood*' superior a 3,84. En total se trata de 482 UL que ocurren juntas 18.999 veces en Econ (el 16,05% de todas las OC) y 750.955 veces en Cumbre (el 3,87% de las OC). Consideramos a estas 482 UL el léxico subeconómico del corpus Econ, ya que parece que está al servicio del léxico económico.

Después de haber encontrado un método para determinar qué léxico general es típico de una lengua especializada – en este caso la económica -, está por resolver la cuestión de encontrar un método para detectar el vocabulario general básico. Para 1.856 UL en Econ el resultado obtenido con el '*log-likelihood ratio*' no se puede considerar estadísticamente significativo. Esto significa que son UL iguales de frecuentes en los dos corpus. Ahora bien, para poder contar estas UL como pertenecientes al '*core vocabulary*' de la lengua española, es necesario ordenarlas por su frecuencia corregida y examinar a partir de qué punto cesa el progreso sustancial de cobertura en el corpus. Este punto se introduce a partir de las 1.025 UL, cuando la progresión de la tasa de cobertura cae por debajo del umbral del 3,5% propuesto por Hazenberg**.**

Finalmente, la aplicación del '*log-likelihood ratio*' a las UL generales de Econ indica 318 UL como significativamente más frecuentes en Cumbre que en Econ. No obstante, al mirar estas UL, muchas de ellas – las más frecuentes – parecen intuitivamente fundamentales para la lengua en general, como se puede constatar en la siguiente lista que presenta las 10 primeras ordenadas según su frecuencia corregida : *hacer, decir, sólo, dar, saber, sí, quedar, menos, vez* y *crear.* De ahí que se haya aplicado una vez más el criterio propuesto por Hazenberg, por lo cual quedan seleccionadas las 180 UL más frecuentes y mejor repartidas de este grupo.

4. Conclusión

La eficacia de las técnicas utilizadas ha sido puesta a prueba de la siguiente manera. Respecto de la tasa de cobertura necesaria para la comprensión de un texto, los resultados de la investigación experimental (por ejemplo Laufer, 1992 ; Nation, 1990, 2001) sugieren que un estudiante medio necesita entender entre el 95% y el 98% de las OC en un texto escrito para que su comprensión de dicho texto sea aceptable, o sea para que obtenga una nota de aprobado. Hazenberg (1994) añade que a partir de un 95% de comprensión, el alumno es capaz de inferir el significado del léxico desconocido en base al contexto.

Ahora bien, juntas, las OC de las UL seleccionadas cubren el 98,22% del corpus Econ, una tasa que supera con creces el 95% requerido según la bibliografía para asegurar una comprensión mínima e, incluso, el umbral del 98% que garantiza una lectura fluida y agradable sin asistencia necesaria. Este resultado se restringe, por supuesto, al corpus examinado. A fin de componer el léxico que cubra por lo menos el 95%, o mejor aun, el 98% de cualquier texto de discurso económico empresarial, se necesita ampliar el presente corpus con otros textos más.

Referencias

Ahmad K. y Rogers M. (2001). Corpus Linguistics and Terminology Extraction. In Wright S. y Budin G. (Eds), *Handbook of Terminology Management,* vol. (2). John Benjamins Publishing Company : 725-760.

Cabré M.T. (1999). *La terminología. Representación y comunicación*. IULA.

Clijsters W. (1990). *Mille lettres d'affaires en chiffres*. Champion/Slatkine.

Dollerup C. *et al*. (1989). Vocabularies in the reading process. *AILA Review*, vol. (6) : 1-33.

Hazenberg S. (1994). *Een keur van woorden.* Ridderprint.

Juilland A. y Chang-Rodríguez E. (1964). *Frequency dictionary of Spanish words.* Mouton&Co.

Kocourek R. (1982). *La langue française de la technique et de la science*. Oscar Brandstetter Verlag.

Laufer B. (1992). How much lexis is necessary for reading comprehension? In Béjoint H. y Arnaud P. (Eds), *Vocabulary and Applied Linguistics*. MacMillan : 126-132.

Müller Ch. (1977). *Principes et méthodes de statistique lexicale*. Hachette.

Nation I.S.P. (2001). *Learning vocabulary in another language.* Cambridge University Press.

Nation I.S.P. (1990). *Teaching and learning vocabulary.* Newbury House.

Pérez Gorostegui E. (1997). *Introducción a la administración de empresas.*Editorial Centro de Estudios Ramón Areces

Rayson P. y Garside R. (2000). Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38 annual meeting of the Association for Computational Linguistics (ACL 2000), 1-8 October 2000, Hong Kong, 1-6, en línea : http://www.comp.lancs.ac.uk/computing/users/paul/publications/rg_acl2000.pdf

Rosengren I. (1971). The quantitative concept of language and its relation to the structure of frequency dictionaries. *Etudes de Linguistique Appliquée*, vol. (1) : 103-127.

Sager J.C. (1990). *A practical course in terminology processing.* John Benjamins Publishing Company.

Sánchez A. *et al*. (2001). *Gran diccionario de uso del español actual*. SGEL.

Stubbs M. (2001). *Words and phrases. Corpus studies of lexical semantics*. Blackwell Publishers.

Sutarsyah C. *et al*. (1994). How useful is EAP vocabulary for ESP? A corpus-based study. *RELC Journal*, vol. (25/2) : 34-50.

Temmerman R. (2000). *Towards new ways of terminology description. The sociocognitive approach*. John Benjamins Publishing Company.

Vangehuchten L. (2003). *El léxico del discurso económico empresarial : elaboración de una metodología con vistas a su descripción y análisis en ELE*. Tesis inédita.

A functional analysis of the linguistic variation in Flemish spoken commercials

Sofie Van Gijsel, Dirk Geeraerts, Dirk Speelman

KU Leuven – RU Quantitative Lexicology & Variational Linguistics
Blijde Inkomststraat, 21 – 3000 Leuven – Belgium
sofie.vangijsel@arts.kuleuven.ac.be, dirk.speelman@arts.kuleuven.ac.be

Abstract

This article discusses a statistical analysis of the language use of a corpus of spoken commercials, broadcast on Flemish radio and television. Advertising typically attempts to persuade the consumer's public by evoking an identification process with the product promoted. To optimize this identification process, copywriters exploit a number of social trends and evolutions, an important part of which is the linguistic register. In this sense, it can be assumed that the commercials form an ideal corpus to study the linguistic variation –and the linguistic attitudes- in present-day Flanders.

The language use in Flanders is characterized by a stratified continuum in which the informal, substandard register of Dutch, termed '*tussentaal'* (literally: 'in-between language'), takes up an (increasingly) central position, between the regional dialects and standard Dutch. This research will focus on the presence of *tussentaal* characteristics in the corpus. To explore the functional motivations which account for the register variation encountered, a regression analysis is carried out, using a number of dependent variables such as the medium, the format and the target group of the commercial.

Keywords: language variation, regression analysis, spoken advertising language, Dutch.

1. Introduction

In this article, the results of a linguistic analysis of a sample of television and radio commercials, broadcast on Flemish channels, are discussed. As will be explained, one of the essential goals of advertising is getting the potential consumer to identify with the advertising message and the product presented. An efficient technique to stimulate this identification process is the use of an appropriate linguistic register. This entails that the corpus chosen provides trend-sensitive linguistic material, offering insights in the language situation in present-day Flanders.

The Flemish language situation is characterized by a stratified continuum of linguistic registers: although a Flemish version of standard Dutch exists, a substandard, yet not dialectal linguistic variant, dubbed '*tussentaal'* (literally 'in-between language'), is often used, especially in less formal situations. The analysis in this article will especially focus on *tussentaal*, recording (the frequencies of) the different linguistic characteristics in the corpus. The correlations between the variation encountered and a number of extralinguistic factors (e.g. the spot format) will be studied in a regression analysis. Thus, this statistical analysis offers insight in the *functional* motivations which account for the *register variation* in the commercials.

The organization of this article is as follows. First, a short introduction to the field of advertising is given, showing the importance of the linguistic register in commercials. A brief

overview of advertising language research will point at the differences with the proposed analysis. The following section discusses the current language situation in Flanders in some more detail, focusing on the position of *tussentaal*. In the next section, the material and method are presented, followed by the analysis and a discussion of the results. Finally, a brief conclusion is given.

2. The language of advertising

Advertising forms part of the 'marketing (communication) mix' (Borden, 1967), a number of techniques aiming at an optimal communication between marketers and the target public. In a number of ways, advertising differs from a prototypical communication situation (Batra *et al.*, 1996): not only is the coding process complicated, since the sender is a group of *spokes-persons* rather than a single person, but the decoding process is also exceptional in that the receiver, who is part of the consumer public, is unable to provide (direct) feedback. As a consequence, in order to create efficient advertising, marketing research aims to analyze how the consumer *processes* commercials. The efficiency of a commercial is enhanced if the consumer *identifies* with the advertising message, since this creates a positive attitude towards the product advertised. This identification process can be brought about by developing a good product *image* (or *brand personality*) or by creating *group norms*, which suggest that certain people or spokespersons, with whom the target public identifies, positively validate the product. Fauconnier and Van der Meiden (1993: 21) point out that in this process, a number of cultural and social factors play a role, an important one of which is the *linguistic register* used.

Broadly speaking, the language of advertising has been analyzed from two perspectives. Some research concentrates on the persuasive –and often misleading- effects of advertising. Geis (1982) for example shows how advertising makes use of techniques such as conversational and pragmatic implicatures, ellipses and hedges, which make the consumer assume stronger claims than are acceptable from a strictly logical point of view. In the same vein, Vestergaard and Schröder (1985) analyze ads in magazines, discussing a number of (rhetoric) devices which exploit cultural presuppositions and images in order to 'lure' the public into purchasing the presented products. Similarly, Goddard (1998) and Cook (2001) study the discourse of advertising, focusing on the interaction between textual (e.g. connotations) and contextual elements (e.g. the music and/or the pictures or the 'narrative voice' of the speaker).

Another strand of research studies the advertising genre from a more linguistic-technical point of view, mostly focusing on commercials in a single language. For example, research analyzing Dutch advertising language includes Lemmens (1994) and Gijsbers *et al.* (1998), who study the use of English in Dutch commercials, while Stroobants (1997) studies the linguistic differences between commercials broadcast in The Netherlands and in Flanders (e.g. the word order of the verbal group).

In this research, the aim is to combine the two perspectives. In an empirical, corpus-based analysis, the linguistic characteristics of the commercials will be analyzed. The variation encountered will be related to a number of extralinguistic factors which *functionally motivate* the register chosen in the advertisement corpus.

3. The present language situation in Flanders

After a century-long struggle for a Dutch standard language in Flanders (see for example Van den Toorn *et al.*, 1997), a Flemish norm for standard Dutch has emerged,[1] which differs from standard Dutch as spoken in The Netherlands in a number of phonetic, morphological, lexicological, and grammatical aspects (see Geeraerts *et al.*, 1999 for an overview). Yet, in more informal strata of Flemish Dutch, the standardization process seems to have halted, giving way to the use of an intermediate, substandard register of Dutch, dubbed *'tussentaal'* (literally '*in-between* language') or '*Verkavelingsvlaams'* (Van Istendael, 1989). This register, which is positioned on the linguistic continuum between the regional dialects and standard Dutch, seems to be a more 'natural' language for many Flemish, while the standard language is often considered an external, artificial linguistic norm (e.g. Jaspaert, 1986; Geeraerts, 1998). Although *tussentaal* functions *de facto* as an informal vernacular on a supraregional level, it is characterized by a number of *substandard* features, especially on the level of pronunciation and morphology. Geeraerts *et al.* (1999) have also shown that the distance between *tussentaal* and Belgian standard Dutch on the one hand is larger than that between the Netherlandic informal register and standard language on the other hand.

It is largely due to its rather recent emergence and substandard character that relatively little research has focused on the (possibly) systematic linguistic characteristics of *tussentaal,* although Smedts (2000) and Geeraerts (2001) present an overview of *tussentaal*-features. In the analysis presented here, the substandard linguistic features occurring in the supraregional corpus of spoken commercials will be identified (see 4.2.2 for an enumeration of the different *tussentaal*-characteristics analyzed). To explore what functional motivations account for the use of *tussentaal*, the linguistic register is correlated to a number of extralinguistic factors.

4. Material and method

4.1. *Material*

The sample of commercials analyzed consists of 150 radio and 150 television ads, broadcast on Flemish, supraregional channels.[2] All material was collected during the Fall of 2000. In order to collect a representative sample, commercials from the different Flemish channels and for different times of the day were recorded. Every commercial was included only once in the sample.

4.2. *Method*

Based on research as presented in 3., the linguistic characteristics encountered in the corpus are identified as either *tussentaal* or standard language. Next, the effect of a number of extralinguistic, independent factors on the language used in commercials is analyzed in a regression analysis. In Section 5, this statistical analysis will be discussed; here, the independent and dependent variables of the statistical analysis are introduced.

[1] This standard language is often dubbed '*VRT-Dutch*', after the national television broadcast channel, VRT, whose news reporters are generally perceived as providing a good example of 'standard Dutch' as it is spoken in Flanders.

[2] For radio ads, material was taped from the Radio 1 and Radio 2, Studio Brussel, Donna and Klara, while for the television commercials, the national channels *VRT* and *Canvas* and the commercial channels *VTM*, *Ka2* and *VT4* were taken into account.

4.2.1. *The independent variables*

A first independent variable which is taken into account is the *medium* of the ad: it can be expected that the linguistic register of the commercial depends on whether it is broadcast on radio or on television. While television ads make use of a number of visual aids to convince the potential consumer, the techniques available for radio commercials are more limited (see also Hagerman, 1990: 159). Therefore, for radio ads, a more creative use of the linguistic register might be expected.

Secondly, the *target public* was a potentially significant factor: is it possible to find register differences in commercials specifically aiming at a public of children (*Child*), adolescents (*Adolesc*) or a more general (mostly adult) public (*Gen*)?[3]

A third factor analyzed is the *spot format*. As Hagerman (1990: 125) explains, a format is 'a combination of structure and content that acts as a catalyst to hold the listener's interest'. Although in works on advertising techniques, the exact number and types of formats are slightly different (see for example Baldwin, 1989 or Batra *et al.*, 1996), for this research, it was sufficient to work with relatively general and easily discernable categories. Broadly speaking, two main types of formats can be distinguished. On the one hand, a number of formats adopt an informative, *rational-argumentative* approach, by focusing on the product itself (so-called *product presentations*) or by showing its use (*demonstration*-spots). Other formats, on the other hand, are more *emotional-suggestive,* in that they attempt to persuade the public by evoking (positive) feelings. An important format type belonging to this second category is the *minidrama*, in which actors play a certain scene (for example, a *slice-of-life* minidrama for cereals could show a family having breakfast). Other formats include *personality* spots, where a character or (cartoon) figure is linked to the product (e.g. *Tony the Tiger* for *Frosties* of Kellog's) and *testimonials*, in which an expert, 'the man on the street' or a celebrity testifies about the product. Minidrama's differ from the other emotional-suggestive formats in that most often, more than one actor is involved, interacting in dialogic speech. Since this dialogic aspect might have an influence on the linguistic register, this spot format is analyzed separately. Finally, each of the different formats discussed can be followed by a *pay off*, a short format element which is typically read by a *voice-over*, who repeats the brand name and/or a slogan. Since it is interesting to test if the linguistic register of the voice-over is different from the language in the spot body, the *format elements* rather than the entire spots are taken as observations in the regression analysis. Thus, each of the format elements of the corpus was assigned to one of the following categories: 1) rational-argumentative spots (*Rat*), 2) minidrama (*Md*), 3) emotional-suggestive spots (*Em*) and 4) pay off *(Po)*.[4]

A last factor analyzed was the *spot subject*, according to 7 categories (toys, technology & communication, cars, financial services, media & culture, products for health and personal care and non-profit ads[5]). Since none of the spot subjects showed a significant relation with the linguistic register, this factor will not be discussed in detail.

[3] Since occasionally, it was rather difficult to draw the line between 'adolescent' and 'adult' spots, only those ads which could clearly be distinguished (for example based on the spokespersons or the product advertised) were classified as adolescent or *Adolesc*-spots.

[4] Of course, the division between the different format types is not always clear-cut; indeed, efficient commercials will often combine different persuasion techniques. Yet, only in very few cases, it was doubtful which format type an element had to be assigned to.

[5] Non-profit spots are sponsored by the government or non-profit organisations (e.g. for a drink-drive campaign) and most often broadcasted on national channels. Therefore, they could be expected to have a higher use

4.2.2. *The dependent variable*

As was mentioned before, the analysis presented focuses on the stratified continuum of spoken language use in Flanders. Thus, a number of spot elements in the corpus, which made use of linguistic registers falling outside this continuum, were consequently excluded from the dataset used in the regression analysis. These were English-spoken elements, a small number of television spots in which no spoken language was used, one purely dialectal spot element and a few spot elements in which Netherlandic Dutch was spoken. Globally speaking, 409 spot elements were analyzed, while 53 were left out.

Instead of working with a binary classification of the spot elements in *tussentaal* and standard language, a *tussentaal*-index was calculated for each of these 409 elements, based on the relative frequency of all *tussentaal*-characteristics vs. all linguistic characteristics encountered in the spot element. This way, the intrinsic continuum of the linguistic register is taken into account. The 17 linguistic characteristics analyzed are the following:[6]

Phonological characteristics:

- word-final t-deletion in words like 'goed' ('good'), 'niet' ('not'), 'wat' ('wat') and 'dat' ('that'), 'met' ('with') and end deletion in 'maar' ('but')
- regional accent of Brabant (the central province of Belgium)
- word-initial h-dropping

Morphological characteristics:

- variants of the personal pronouns of the 1st person singular (*(e)kik* instead of 'ik'), the 2nd person singular (*gij* instead of 'jij'/'you' and the 3rd person singular (*hem* instead of 'hij'/'he')
- variants of the attributive adjective when followed by a *d, t, b, h* or vowel
- variant of definite article 'de' ('the'): *den* or *nen*
- variant of the indefinite article 'een' ('a'): *ne* or *e*
- variants of the demonstrative pronoun (*dieje/diene/dezen* for 'die/deze' or 'these/those'), possessive pronoun (e.g. 'mijn*en*' for 'mijn'/'mine') and *een/geen* ('one/none'), if followed by a *d, t, b, h* or vowel
- variant of the diminutive: *–(s)ke* instead of '–je' (and its allomorphs)
- variant of imperative singular: *root+t* instead of *root*

5. Linear regression

As mentioned before, a *linear regression* analyzes the impact of the independent factors (in this case medium, format type, target group and subject) on the dependent factor, the linguistic register used in the spot elements.[7] The statistical model is built using a *forward stepwise procedure*: the independent variables are added to the model in their order of significance; if

of the standard language. Although a slightly higher use of standard Dutch was found, the result was not significant.

[6] Although we find a few other morphological variants (such as plural pronouns), some minor *syntactic* deviations (e.g. in the order of the verbal group or the double negation) and *lexical* differences (e.g. substantives and interjections), these elements were extremely infrequent in the corpus. Therefore, they were not analyzed in further detail.

[7] In *Appendix A*, a table is included which gives an overview of the absolute numbers of the independent variables in the entire set of 409 observations, split up for '*tussentaal'* and 'standard language'. For ease of representation, in this table, each spot element containing one or more *tussentaal*-characteristics was classified as '*tussentaal'*. Of course, it should be kept in mind that, since the linguistic registers form a continuum, the boundary between the *tussentaal* and standard Dutch is rather artificial; in the regression analysis, a *tussentaal-index* is used rather than a binary classification to account for this gradualism.

the next variable to be added to the model does not contribute significantly to the proportion of explained variation, the analysis automatically stops. Not only will it be possible to test if all the extralinguistic factors analyzed have a statistically significant effect on the linguistic register, but also, the statistical output indicates the *effect of the individual variables* by way of the *regression coefficient*. This coefficient, also called the *B-value* or *estimate* of the independent variable, gives a measure of how the dependent variable (in our case, the *tussentaal*-index) increases if the other variables in the model are kept constant. Further, it is possible to evaluate the *global model* by interpreting the *R-squared value*: this value between 0 and 1 indicates how much of the variation in the data is explained by the model. Finally, a *p-value* for the significance of the global model is also given.

As was mentioned before, the results for 'subject' were not significant. Therefore, it comes as no surprise that the stepwise procedure does not include this factor in the model. The other three factors significantly contribute to the explained variation, with 'format' being added first to the model, followed by 'target group' and 'medium'. An overview of the p-values of the different factors, and the estimates (or B-values) with their confidence intervals (C.I.) is given in *Table 1*:[8]

	p-value	*Estimate*	*C.I.*
Format: minidrama	< 0.001	0.312	0.270-0.352
Public: adolescents	< 0.001	0.225	0.118-0.331
Medium: television	0.0125	-0.044	-0.079-(-0.009)

Table 1. The P-values, estimates and respective confidence intervals for the significant independent variables influencing the tussentaal-*index of the spot elements*

It is clear that the factor 'minidrama' is highly significant ($p < 0.001$), with the estimate indicating that the model gives this format element 31.2% more for the *tussentaal*-index than to the other format elements. For the factor 'target public', the *estimate* is 0.225 for the value 'adolescents', which shows that *tussentaal* is predicted to be 22.5 %higher in *adolescent*-spot elements than in those having children or adults as main target audience. Admittedly, the confidence interval is rather large here ([0.118-0.331]), due to the small number of *Adolesc*-elements in the sample (see Appendix A). The third most important factor influencing the linguistic register is the *medium*, with 'television' predicting the *tussentaal*-index to decrease with 5 %. Here, the confidence interval is much smaller ([-0.079-(-0.009)]), due to the large number of observations (162 television-formatelements).

Next, it is possible to evaluate the *global model*, which takes up 'format element', 'target group' and 'medium' as factors in the regression. The model is highly significant ($p < 0.001$), which means that the chance that the *tussentaal*-index is not influenced by the independent factors (as the null hypothesis would claim) is virtually non-existent. The *R-squared value* is 0.492, which means that the model explains about 50 %of the variation in the data. Therefore, it can be concluded that the model predicts the degree of *tussentaal* in the spotelements fairly well.

[8] The values of the confidence interval indicate that there is a 95% chance that the real value, for which the B-value is an estimate, lies between these values. The confidence interval will typically be larger if the estimate is based on relatively few observations; this should be taken into account when interpreting this measure.

The results show that the use of *tussentaal* in the corpus of Flemish spoken commercials is strongly influenced by a number of independent factors. A first important factor is the *spot format*: in dialogic *minidramas*, *tussentaal* is used extensively. Most likely, this informal register is used to make the spot sound more 'natural' and 'authentic'. The recurrent combination of a *tussentaal*-minidrama with a pay off in which standard Dutch is spoken shows how the *style-shifting* is used consciously and strategically to initiate and mark a situational change (see e.g. Milroy and Gordon, 2003 for a discussion of the notion of 'initiative style-shift'). Secondly, in radio commercials, *tussentaal* is used more often than in television spots. This may well be due to the fact that radio spots can only make use of sound elements to persuade the consumer, while television ads can exploit a wide array of (audio)visual methods to evoke an identification process. Finally, it is important that the *tussentaal*-index is higher in spots aiming at a public of adolescents. As Bell (2001) and Milroy and Gordon (2003: 207) explain, in order to express identification with particular groups (which may well be geographically distant, as is the audience of commercials), speakers can exploit available linguistic resources. Here, the *adolescent*-audience brings about a strategic style-shift towards a less standardized register. Therefore, it might be hypothized that *tussentaal* has young, even somewhat rebellious, connotations, as opposed to the 'conformist' norm of the standard language.

These conclusions tie in with independent linguistic research (e.g. Jaspaert, 1986; Deprez, 1994; Geeraerts, 1998), which shows that many Flemings consider the standard language as an artificial, external norm rather than an authentic or natural register, especially in slightly less formal, spoken language. Also, the connotations which seem to be associated with *tussentaal* are in line with what Geeraerts (2001) shows for the language used in Flemish soaps: *tussentaal* is often used when a natural, informal register is required to ensure easy identification by the spectators.[9] It seems to be the case that in the advertisement corpus analyzed, the 'authentic', natural-sounding register *tussentaal* is equally exploited as an efficient technique to enhance an identification process of the consumer with the product.

6. Conclusion

In this article, the analysis of a corpus of spoken commercials broadcast on Flemish radio and television was discussed. The advertisements provided an interesting view on the current language situation in Flanders, which is characterized by a stratified continuum of linguistic variants. More specifically, the use of the 'in-between' register *tussentaal*, filling the gap between 'dialects' and 'standard Dutch', was studied in detail. In order to find out what *functional* motivations lie behind the language variation used in the spots, a regression analysis was carried out, investigating the influence of four independent factors on the linguistic register. The results show that radio (as opposed to television) spots, an adolescent target public and dialogic minidramas significantly increase the *tussentaal*-index. These spot types all attempt to sound natural, informal and realistic; connotations which typically seem to be associated with the register *tussentaal*.

[9] Binnemans (2003) comes to similar conclusions in her analysis of the Flemish television travel program '2x Enkel': while the young, somewhat 'wild' male reporter of the program uses *tussentaal* quite intensively, in order to sound natural, casual and a bit rebellious, the older, female reporter, who often presents the more informative sections of the program, attempts to use a more formal, standardized linguistic register.

References

Baldwin H. (1989). *How to Create Effective TV Commercials.* NTC Business Books.

Batra R., Myers J.G. and Aaker D.A. (1996). *Advertising Management.* Prentice-Hall International.

Bell A. (1999). Styling the other to define the self. *Journal of Sociolinguistics*, vol. (4): 523-541.

Binnemans T. (2003). *Taal en Identiteit: Een Functionele Analyse van de Taalvariatie in het Televisieprogramma* 2x Enkel. Unpublished dissertation, University of Leuven.

Cook G. (2001). *The Discourse of Advertising (2nd edition)*. Routledge.

Deprez K. (1994). Waarom ze hun Vlaams liever zien verdwijnen. *Onze Taal,* vol. (62): 251-253.

Fauconnier G. and Van der Meiden A. (1993). *Reclame. Een Andere Kijk op een Merkwaardig Maatschappelijk Fenomeen*. Dick Coutinho.

Geeraerts D. (1998). VRT-Nederlands en Soap-Vlaams. *Nederlands van Nu*, vol. (3): 75-77.

Geeraerts D., Grondelaers S. and Speelman D. (1999). *Convergentie en Divergentie in de Nederlandse Woordenschat. Een Onderzoek naar Kleding- en Voetbaltermen*. Meertens Instituut.

Geeraerts D. (2001). Everyday language in the media. The case of Belgian Dutch soap series. In Kammerer M., Konerding K., Lehr A., Storrer A., Thimm C. and Wolski W. (Eds), *Sprache im Alltag. Beitrge zu neuen Perspektiven in der Linguistik Herbert Ernst Wiegand zum 65. Geburtstag gewidmet*. Walter de Gruyter: 281-291.

Geis M. (1982). *The Language of Advertising.* Academic Press.

Gijsbers I., Gerritsen M., Korzilius H. and van Meurs F. (1998). Engels in Nederlandse tv-reclame. *Onze Taal*, vol. (7/8): 175-177.

Goddard A. (1998). *The Language of Advertising: Written Texts*. Routledge.

Goossens J. (2000). De toekomst van het Nederlands in Vlaanderen. *Ons Erfdeel*, vol. (43): 3-13.

Hagerman W.L. (1990). *Broadcast Advertising Copywriting.* Stoneham. Butterworth Publishers.

Jaspaert K. (1986). *Statuut en Structuur van Standaardtalig Vlaanderen*. Leuven Universitaire Pers.

Lemmens M. (1994). Amazing…moderne reclametaal. *Onze Taal*, vol. (7/8): 176-177.

Milroy L. and Gordon M. (2003). *Sociolinguistics: Method and Interpretation*. Blackwell.

Rietveld T. and van Hout R. (1993). *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter.

Smedts W. (2000). *Project Taalhantering VRT 1999-2000. Eindevaluatie.* VRT-project Taalhantering, Internal Report, University of Leuven.

Stroobants K. (1997). Noord/Zuid-verschillen in de taal van de reclame. *Taal en Tongval*, vol. (10): 140-151.

Van den Toorn M.C. (Ed.) (1997). *Geschiedenis van de Nederlandse Taal*. University Press.

Van de Velde H. (2000). 100 jaar uitspraak. *Nederlands van Nu*, vol. (3): 35-38.

Vestergaard T. and Schrøder K. (1985). *The Language of Advertising.* Basil Blackwell.

Van Istendael G. (1989). *Het Belgisch labyrint*. *De schoonheid der wanstaligheid*. Arbeiderspers.

Appendix

		Standard language (306)			*Tussentaal (103)*			*Total (409)*
		Child	Adolesc	Gen	Child	Adolesc	Gen	
Radio		3	2	164	0	7	71	247
	Rat	2	1	47	0	0	2	52
	Md	0	0	18	0	4	67	89
	Em	1	0	2	0	0	1	4
	Po	0	1	97	0	3	1	102
TV		28	3	106	5	0	20	162
	Rat	17	1	49	0	0	0	67
	Md	3	0	12	4	0	17	36
	Em	0	0	11	1	0	2	14
	Po	8	2	34	0	0	1	45
		31	5	270	5	7	91	**409**

Table 2. The absolute numbers of the independent variables 'target public', 'format' and 'medium' for the 409 format elements, split up for 'standard language' and 'tussentaal'

Legend:

Target public:
- Child: children
- Adolesc: adolescents
- Gen: general

Format:
- Rat: rational-argumentative
- Md: minidrama
- Em: emotional-suggestive
- Po: pay off

Medium:
- Radio
- TV

Polysémie et calcul du sens

Fabienne Venant

LaTTICe- ENS – 1 rue Maurice Arnoux – 92120 Montrouge – France
fabienne.venant@ens.fr

Abstract

Polysemy is a pervasive phenomenon in natural languages, but it remains a vexing issue for natural language computing. In order to deal with this problem, Victorri and Fuchs (1996) proposed the model of *dynamical construction of meaning*. Each polysemic unit is associated to a semantic space. The meaning of the unit in a given sentence corresponds to a more or less restricted area of the semantic space, resulting from a dynamical interaction with all other units of the sentence. Ploux and Victorri (1998) designed a software, called Visusyn, allowing an automatic construction of the semantic space of a polysemic unit. The algorithm is based on the analysis of a large graph of synonyms (www.unicaen.crisco.fr). Visusyn was extended to take into account the data from one corpus (base Frantext categorisée). The performance of Visusyn was compared to that of french speakers in a disambiguation task. The results of this experiment strongly incited us to start a theoretical mathematic and informatic study of the graph of synonyms. This study will be crucial for a better understanding of french lexical organisation.

Résumé

La polysémie est un phénomène omniprésent dans le langage, mais il reste problématique dans le cadre du traitement automatique des langues. Nous nous proposons d'aborder ce problème à l'aide du modèle de *construction dynamique du sens*, proposé par Victorri et Fuchs (1996). On associe à chaque unité polysémique un espace sémantique. Le sens de l'unité dans un énoncé donné correspond à une région plus ou moins étendue de cet espace, déterminée par l'interaction dynamique de toutes les unités présentes dans l'énoncé. Ploux et Victorri (1998) ont développé Visusyn, un logiciel permettant de construire automatiquement l'espace sémantique d'une unité polysémique. L'algorithme repose sur l'analyse d'un grand graphe de synonymie (www.unicaen.crisco.fr). Nous avons étendu Visusyn afin qu'il prenne en compte les données issues d'un corpus (Frantext catégorisée). Nous avons comparé les performances de Visusyn avec celles de locuteurs du français dans une tâche de désambiguïsation. Les résultats de cette expérience nous ont encouragé à entreprendre une étude théorique mathématique et informatique du graphe de synonymie. Celle-ci devrait s'avérer cruciale pour une meilleure compréhension de l'organisation du lexique français.

Mots clés : polysémie, calcul du sens, espace sémantique, désambiguïsation.

1. La polysémie en linguistique computationnelle

La plupart des unités lexicales que nous utilisons ont plusieurs sens. Loin de nous gêner pour communiquer, ce phénomène, appelé polysémie, est au contraire source de richesse et de souplesse dans les langues. Nous sommes habitués à manier les indices contextuels et nous comprenons instantanément le sens de n'importe quel mot polysémique dans n'importe quel énoncé. Pourtant dès que l'on veut automatiser une telle performance, la polysémie devient un véritable problème et elle donne bien du souci aux chercheurs en traitement automatique du langage. La prise en compte de la polysémie en TAL se traduit par la question suivante : « comment associer automatiquement un sens à un mot dans un énoncé donné ? ». La tâche s'effectue donc en deux étapes : d'abord déterminer tous les sens possibles pour chaque mot susceptible d'être désambiguïsé et ensuite déterminer quel sens est le bon en contexte. Les ordinateurs vont utiliser les mêmes indices que nous, à savoir le contexte. Ce qui leur manque

c'est toute notre connaissance du lexique et de son organisation. C'est là tout l'enjeu des tâches de désambiguïsation. Il y a deux pistes suivies majoritairement : l'une consiste à utiliser les distinctions de sens présentes dans les dictionnaires, l'autre à utiliser des méthodes statistiques pour repérer des patterns de cooccurrences des mots en contexte. En 1998, *Computational Linguistics* a spécialement consacré un article à la question de la désambiguïsation sémantique (Ide et Véronis, 1998).

Dans les années 80, les ressources lexicales à grande échelle (dictionnaires électroniques, glossaires, thésaurus, ontologies…) se sont développées et beaucoup de travaux ont utilisé les divisions de sens fournies par ces outils. L'idée est que le sens le plus probable pour une occurrence d'un mot donné est celui qui va maximiser une certaine relation d'affinité avec le contexte de cette occurrence. Lesk (1986) a créé une méthode permettant de relier des définitions si elles ont des mots en commun. La désambiguïsation d'un mot en contexte se fait en choisissant pour lui et les mots qui l'entourent les définitions qui se recoupent le plus. Cette méthode est très sensible à la présence ou non d'un mot dans une définition et pose problème en cas de définitions lapidaires. D'autres études ont donc cherché à l'améliorer en utilisant d'autres indices (Walker, 1987 ; Guthrie *et al.*, 1991 ; Wilks *et al.*, 1990). Véronis et Ide (1990) ont prolongé la méthode de Lesk en créant un réseau de neurones à partir des définitions du Collins English Dictionary. Wilks *et al.* (1993) ont beaucoup réfléchi à la façon d'utiliser de façon optimale les ressources électroniques pour identifier les sens des mots polysémiques.

Plus récemment on a vu se développer des méthodes de désambiguïsation sémantique sur corpus. Il s'agit d'analyser les mots qui cooccurrent avec les mots polysémiques sur des corpus à grande échelle. Ces systèmes s'entraînent à modéliser le sens de chaque mot, en fonction de leur contexte, à partir de corpus d'exemples sémantiquement étiquetés (de 50 à 100 phrases). Ils choisissent ensuite le sens le plus adéquat pour une nouvelle occurrence d'un mot dans le texte à traiter. L'adéquation d'un sens est calculée à partir d'une mesure de similarité entre les caractéristiques des sens modélisés et celles du contexte de l'occurrence considérée. Une des difficultés de la tâche vient de l'inventaire des sens lui-même. La plupart des travaux réalisés reposent sur des dictionnaires traditionnels, ou des ressources électroniques comme Wordnet qui ne diffèrent pas énormément en termes de division de sens. Le problème est que les dictionnaires ont été réalisés pour un usage humain et non pas automatique. Ils manquent donc d'informations pragmatiques utiles à la désambiguïsation. D'autre part, l'inconsistance des dictionnaires est bien connue des lexicographes (Kilgariff, 1994). Véronis (2001) pense qu'on ne pourra pas progresser en désambiguïsation sémantique tant que les dictionnaires n'incluront pas dans leurs définitions des critères distributionnels ou des indices de surface (syntaxes, collocations,…). C'est pourquoi au sein de son équipe, Reymond travaille à la réalisation d'un dictionnaire « distributionnel » spécialement adapté au problème de la désambiguïsation par des machines (Reymond, 2001). Il s'agit d'organiser les mots en lexies possédant des propriétés distributionnelles cohérentes. Audibert travaille, à partir de ce dictionnaire, à étudier les différents critères de désambiguïsation (cooccurrence, n-grammes, information sur le domaine, synonymes des mots en cooccurrence,...) (Audibert, 2002-2003).

Une autre voie de recherche suivie par Véronis (2003), toujours dans l'idée de pallier aux insuffisances des dictionnaires classiques en matière de discrimination des sens, est l'utilisation d'un graphe de cooccurrences. Il s'agit de déterminer automatiquement les différents usages d'un mot dans une base textuelle (en l'occurrence google). L'algorithme est basé sur la recherche des zones de forte densité du graphe de cooccurrences et permet contrairement aux méthodes classiques d'analyse textuelle (comme les vecteurs de mots), d'isoler des

usages très peu fréquents. Jean Véronis met ici en application le conseil de Wittgenstein : « Don't look for the meaning, but for the use. »

Une autre solution est de travailler sur des relations paradigmatiques entre mots (synonymie, antonymie,…). Comme le remarquent Edmond et Hirst (2002) un mot peut exprimer une myriade d'implications, de connotations en plus de son sens dans les dictionnaires. Un mot a des synonymes (il s'agit ici de relation de synonymie partielle) qui diffèrent de lui par ces nuances de sens. Ils cherchent à développer un modèle computationnel de la connaissance lexicale qui rende compte adéquatement de la « presque synonymie » et qui dans une tâche de traduction automatique puisse choisir le bon mot, celui qui va rendre compte de la nuance de sens exacte, dans un contexte donné. L'idée étant de pouvoir rendre compte des sens indirects, flous ou dépendant du contexte ignorés des systèmes actuels.

À la lumière de tous ces travaux, on peut s'étonner avec Véronis (2001) du fait que la pertinence cognitive ne soit jamais recherchée. Une expérience qu'il a menée montre que les humains eux-mêmes ont de piètres performances quand il s'agit d'associer un sens d'un dictionnaire à une occurrence d'un mot dans un énoncé. Mis à part Edmond et Hirst, on s'interroge peu sur le fait qu'une occurrence d'un mot puisse jouer sur plusieurs sens possibles sans qu'on puisse trancher entre les deux. Ce phénomène qu'on appelle indétermination est pourtant au cœur même de l'expressivité d'une langue. Enfin, aucune des méthodes actuelles ne s'interroge réellement sur l'organisation du lexique. Même les méthodes basées sur des calculs de similarités ne cherchent pas à représenter les distances sémantiques entre sens et ne parviennent pas à organiser correctement les sens obtenus.

Le modèle que nous utilisons s'appuie sur un espace sémantique où sont organisés les différents sens d'un mot. La polysémie est le mécanisme central dont nous cherchons à rendre compte. Le calcul du sens d'un énoncé est un processus dynamique au cours duquel les sens des différents mots s'influencent mutuellement et qui aboutit simultanément à la détermination du sens de chacun des mots et à un sens global pour la phrase. C'est le principe de la compositionnalité gestaltiste défini par Victorri et Fuchs (1996). Ce modèle a été implémenté en utilisant deux ressources lexicales : un dictionnaire électronique de synonymes et un grand corpus.

2. Calcul dynamique du sens

Le principe de compositionnalité gestaltiste implique une modélisation dans le cadre des systèmes dynamiques. Cela permet d'éviter le cercle vicieux du au fait que la plupart des unités sont polysémiques, et que pour calculer le sens de chacune d'elles on a besoin de connaître les sens des autres, et réciproquement.

La donnée d'une dynamique sur un espace revient à spécifier les contraintes qui s'exercent en chaque point de cet espace et permet d'obtenir les points de stabilisation qui correspondent aux solutions du problème. Il s'agit donc ici d'associer à chaque unité linguistique un espace, appelé espace sémantique, muni d'une structure mathématique précise où le sens de l'unité dans chacun de ses emplois est représenté par une région de l'espace.

Les unités cotextuelles définissent une « fonction potentielle » sur l'espace sémantique. Les valeurs du potentiel inférieures à un certain seuil déterminent une région de l'espace sémantique qui représente le sens de l'unité dans l'énoncé considéré.

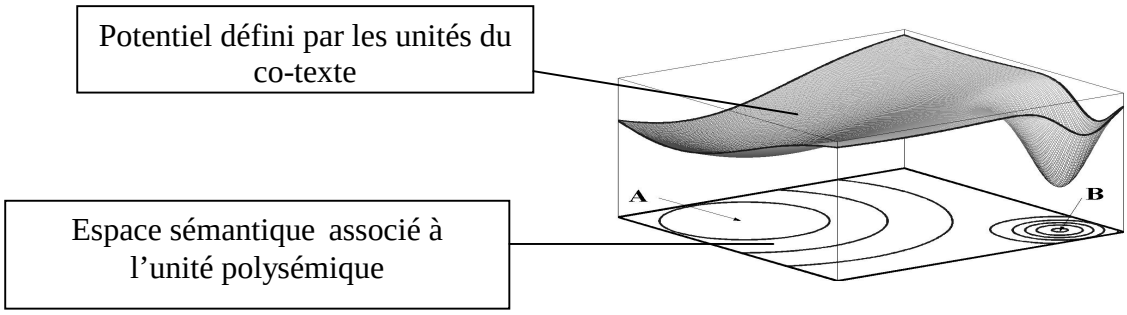


Figure 1. Représentation d'une fonction potentielle sur un espace sémantique bidimensionnel

3. Visusyn

Pour construire et représenter de façon automatique l'espace sémantique associé à un mot, nous utilisons le logiciel VISUSYN développé par Ploux et Victorri (1998). Pour déterminer automatiquement les paramètres de l'espace sémantique, Visusyn utilise un algorithme basé sur l'analyse d'un graphe de synonymie. Ce graphe nous est fourni par le Dictionnaire Electronique des Synonymes (D.E.S.) du laboratoire CRISCO (www.crisco.unicaen.fr).

Prenons un exemple : on s'intéresse ici à l'adjectif *sec*. Le D.E.S. fournit la liste des synonymes de *sec* (au nombre de 63). Il en construit le graphe : les sommets sont *sec* et ses synonymes, deux unités sont en relation si elles sont synonymes. On trouvera en figure 2 un extrait du graphe de synonymie associé à *sec*.

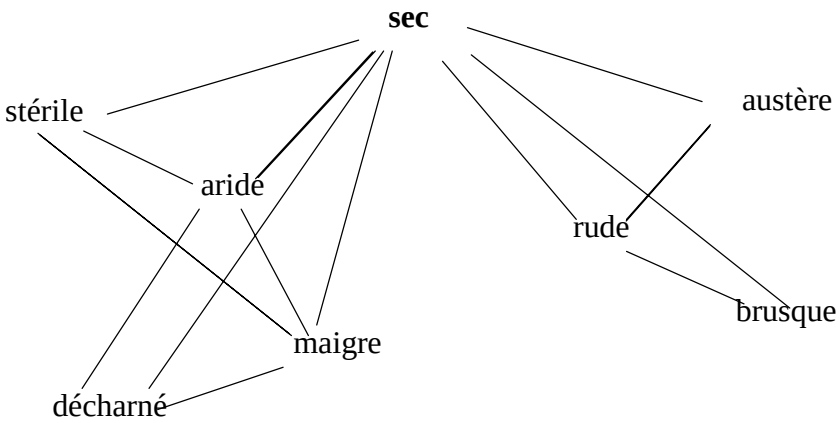


Figure 2. Un extrait du graphe de synonymie de sec

L'idée sous jacente à la construction de l'espace sémantique est semblable à celle développée par Edmond et Hirst (2002). Un simple synonyme n'est généralement pas suffisant pour définir un sens précis d'une unité. On voit ici que *aride* est à la fois synonyme de *décharné* et de *stérile*, ce qui correspond à deux sens distincts de *sec*, l'un lié à la maigreur l'autre au caractère improductif. Or les points de notre espace sémantique doivent correspondre à des sens précis de l'unité. C'est pourquoi nous avons recours à la notion de clique. Une clique est un sous-graphe complet maximal, c'est à dire un ensemble de sommets, le plus grand possible, reliés deux à deux. La portion de graphe que nous avons représentée ici contient 4 cliques : *Aride-décharné-maigre-sec* ; *Aride-maigre-stérile-sec* ; *Austère-rude-sec* et *Brusque-rude-sec*. Chaque clique correspond à une nuance possible de sens pour *sec*.

Le D.E.S. fournit la liste des cliques du graphe correspondant à une unité donnée et ce sont ces cliques que nous allons représenter. Notre espace sémantique est une projection en deux dimensions du nuage formé par les cliques dans l'espace multidimensionnel engendré par les synonymes de l'unité lexicale considérée. Il est muni de la métrique du χ^2. C'est elle qui s'est en effet avérée la plus efficace pour obtenir une représentation respectant la notion intuitive de proximité entre sens. On trouvera une représentation de l'espace sémantique de *sec* en figure 3.

4. Désambiguïsation d'un adjectif

Outre la construction de l'espace sémantique associé à une unité, Visusyn permet d'obtenir la visualisation d'une zone associée à chaque synonyme de l'unité étudiée. Nous avons étendu ce logiciel afin qu'ils puissent prendre en compte des données issues d'un corpus pour visualiser la zone dans laquelle un nom contraint un adjectif à prendre son sens. Nous illustrons ici ces deux propriétés dans le cas de l'adjectif *sec* et nous montrons ensuite comment nous les avons utilisées dans une expérience de désambiguïsation de l'adjectif *sec* en rection nominale.

4.1. *Sémantique de l'adjectif sec.*

Sec est un adjectif très polysémique mais dont on peut regrouper les sens en six acceptions principales :

(1) qui manque d'eau : *du sable sec*

(2) maigre, décharné : *un homme grand et sec*

(3) stérile, improductif : *rester sec aux questions du professeur*

(4) qui manque de sensibilité, qui ne se laisse pas attendrir, égoïste : *un cœur sec*

(5) bref, abrupt, qui manque de douceur : *un coup sec*

(6) seul : *un atout sec.*

Bien que ces sens soient très différents, ils peuvent être reliés les uns aux autres par une « ressemblance de famille » à la Wittgenstein. Les sens (1), (2) et (3) se rejoignent lorsque *sec* qualifie de la végétation. De même les sens (3) et (4) sont liés : une personne sèche au sens d'égoïste est quelqu'un de stérile en termes d'empathie et de don de soi. On sent aussi une relation entre le sens (5), qui s'applique à des événements, et le sens (4) qui caractérise un comportement mal dégrossi. Ce sont toutes ces proximités de sens dont notre représentation doit rendre compte et on peut voir sur la figure 3 que le résultat est plutôt satisfaisant.

4.2. *Zone de pertinence associée à un synonyme.*

À chaque synonyme du mot vedette, on associe une fonction dont les bassins représentent de façon plus précise la zone de sens occupée par ce synonyme. Cette fonction permet de visualiser la région de l'espace sémantique dans laquelle la relation de synonymie entre le mot vedette et le synonyme considéré est pertinente. Elle est calculée sur l'ensemble des cliques en donnant un poids égal à 1 aux cliques contenant ce synonyme, et un poids égal à –0.1 aux cliques ne le contenant pas.

Appelons $u_1, u_2, \ldots, u_m$ les synonymes et $c_1, c_2, \ldots, c_c$ les cliques. La valeur de la fonction associée au synonyme u_j au point de coordonnées (x ;y) est donnée par :

$$f_j(x,y) = 1 - \max\left(0.1 - \sum_{k=1}^{c} a(k,j)\, e^{\frac{(x_k - x)^2 - (y_k - y)^2}{\delta^2}}\right)$$

où (x_i, yi) sont les coordonnées du point

représentant la clique c_i dans l'espace sémantique, a(k,j)=1 si u_j appartient à la clique c_k, -0.1 sinon et $\delta = \frac{\max(dx, xy)}{10}$ où dx et dy mesurent les longueurs des intervalles de coordonnées.

À titre d'exemple, on trouvera en figure 4 et 5 les fonctions potentielles des adjectifs *brusque* et *a*

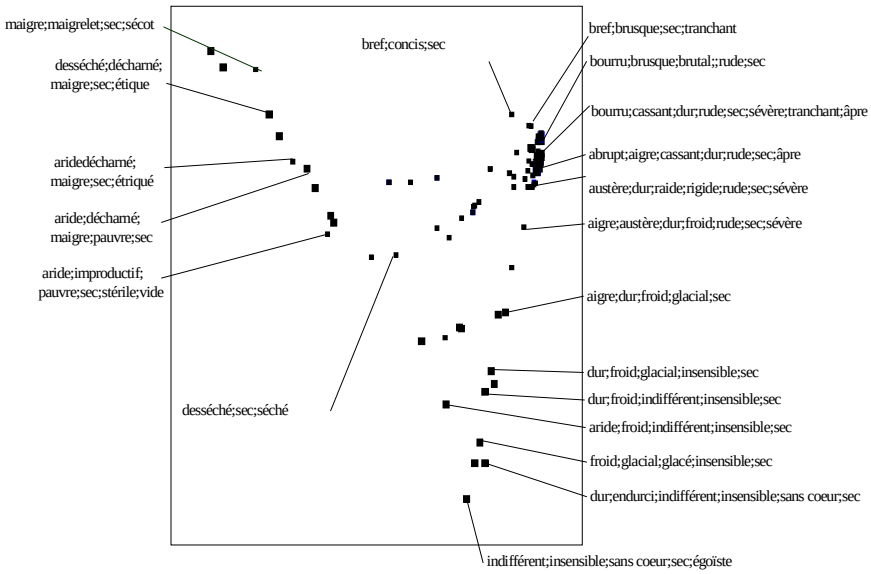


Figure 3. Espace sémantique de l'adjectif sec (63 synonymes, 94 cliques)

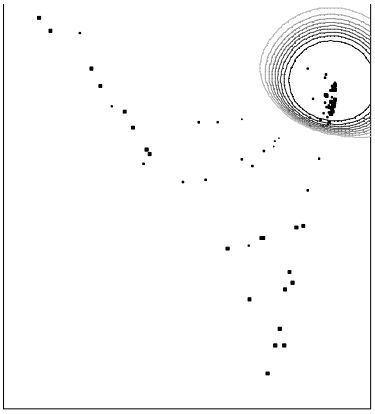


Figure 4.
Fonction potentielle associée à brusque

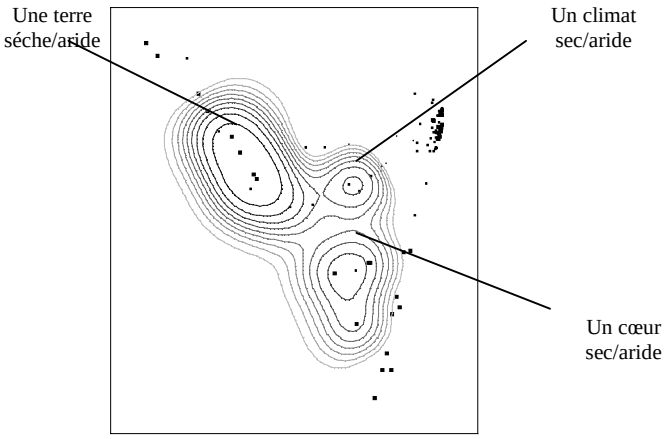


Figure 5.
Fonction potentielle associée à aride

On appelle n_{ij} le nombre réel d'occurrences du couple (e_i,uj) dans le corpus. S'il n'y avait pas d'affinité particulière entre certains noms et certains adjectifs, les couples seraient équidistribués. Le nombre d'occurrences de chaque couple (e_i,uj) ne dépendrait donc que de la fréquence des deux mots pris indépendamment dans le corpus. Appelons m_{ij} ce nombre moyen « théorique ». Alors on a : $m_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ avec $n_{i\bullet} = \sum_{k=1}^{n} n_{ik}$ et $n_{\bullet j} = \sum_{k=1}^{m} n_{kj}$. Pour mesurer

l'affinité d'un nom et d'un adjectif, il faut donc comparer n_{ij} et m_{ij}. Nous définissons ainsi le degré d'affinité d_{ij} du nom e_i avec l'adjectif u_j : $d_{ij} = f(\frac{m_{ij}}{n_{ij}})$ où $f(x) = \frac{x}{2}$ ssi 0<x<2 et $f(x) = 1$ ssi x>2. Le degré d'affinité a_{ik} du nom e_i avec la clique c_k est donné par la formule suivante : $a_{ik} = \frac{\sum_{j=1}^{n} d_{ij} p_{ij} x_{kj}}{\sum_{j=1}^{n} p_{ij} x_{kj}}$ avec $p_{ij} = \frac{m_{ij}}{\sum_{k=1}^{c} x_{kj}}$ et $x_{kj} = 1$ ssi $u_j \in c_k$.

Voici par exemple le nombre de cooccurrences de *coup* avec les adjectifs suivants :

bref : 67, brusque : 48, tranchant : 0, sec : *173, maigre : 0, maigrelet* : *0, sécot : 0.*

et les degrés d'affinité avec deux cliques contenant ces adjectifs : Sec ; bref ; brusque ; tranchant : 90 % et sec ; maigre ; maigrelet ; sécot : 12 %.

On utilise ensuite ce degré d'affinité pour construire une fonction potentielle associée à chaque nom. Cette fonction permet de visualiser la zone de sens pertinente dans le contexte du nom considéré. La valeur de la fonction potentielle associée au nom e_i au point de coordonnées (x ;y) est donnée par le calcul suivant :

$$g_i(x,y) = 1 - \max\left(0, \sum_{k=1}^{c} b(i,k) e^{\frac{(x_k - x)^2 + (y_k - y)^2}{\delta^2}}\right) \text{où } b(i,k) = 2a_{ik} - 0.8 .$$

On peut voir en figure 6 que les résultats obtenus pour des mots comme *fleur* et *ton* sont très bons. La fonction associée à *fleur* possède un unique bassin très étroit. *Sec* a dans ce contexte un sens très précis qui est celui de manque d'eau. Pour *ton*, on obtient un bassin très large qui couvre la presque totalité de la partie droite de l'espace sémantique. On a dans le cas de *ton sec* une indétermination. Un *ton sec* l'est aussi bien d'un point de vue psychologique (en bas à droite de l'espace) que physique (en haut à droite). Il existe cependant des noms comme *lit* et *visage* pour lesquels les potentiels désambiguïsateurs ne correspondent pas aux contraintes réelles qu'ils exercent sur la sémantique de *sec*. C'est pourquoi il nous a paru nécessaire d'étudier plus en détails les résultats obtenus afin de dégager les forces et les faiblesses de la méthode.

Fleur sèche Ton sec

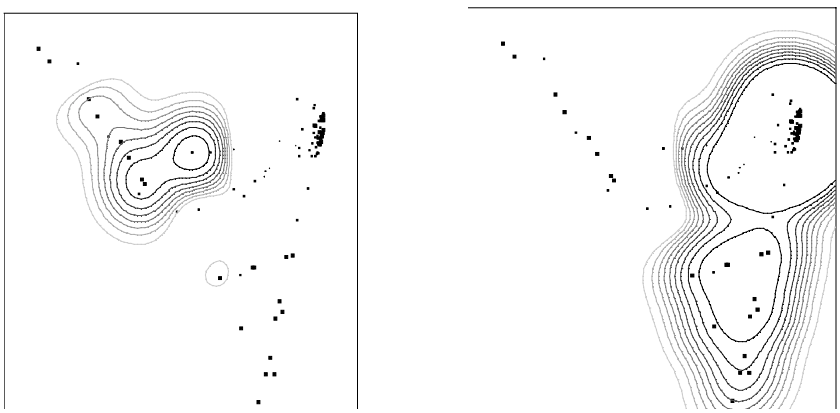


Figure 6. Potentiels désambiguisateurs des noms fleur et ton

4.3. *Évaluation*

Nous avons voulu évaluer la pertinence de nos calculs en comparant les résultats obtenus automatiquement à ceux obtenus par des locuteurs du français. Nous avons donc conçu une tâche de désambiguïsation réalisable à la fois par Visusyn et par des sujets. Il s'agit de sélectionner parmi les 5 synonymes proposés, celui ou ceux qui décrit le mieux le sens de *sec* en présence d'un nom donné. Les synonymes ont été sélectionnés en fonction de leurs indices de similitude. Ces indices sont calculés à partir du graphe de synonymie. L'indice de similitude entre deux mots est égal au nombre de liens qu'ils ont en commun, divisé par le nombre de liens qu'ils ont à eux deux. Il indique donc leur proximité sémantique. On calcule deux indices de similitude entre un synonyme de *sec* et *sec* lui même. Le premier i_1 est obtenu en comptant les liens avec l'ensemble des synonymes de *sec*, l'autre i_2 en restreignant à l'ensemble des synonymes du synonyme considéré. Pour notre travail, nous avions besoin de synonymes de *sec* qui ne soient pas trop ambigus. C'est pourquoi nous avons choisi des synonymes dont la polysémie est plus faible que celle de *sec*, c'est à dire des synonymes pour lesquels i_2 est supérieur à i_1. D'autre part nous avons veillé à ce que ces synonymes correspondent à des cliques bien réparties sur l'espace sémantique. Nous avons ainsi sélectionné les cinq adjectifs : *décharné, desséché, brusque, glacial, stérile.*

Pour réaliser la tâche, Visusyn calcule un taux d'adéquation entre le nom et l'adjectif en mesurant le recouvrement des fonctions potentielles du nom et de l'adjectif. Ce taux est ensuite arrondi et permet d'attribuer une note entre 1 et 5 à l'adjectif. Plus les zones de sens de l'adjectif et du nom sont proches, plus la note est proche de 1. Les sujets doivent eux aussi évaluer l'adéquation entre le synonyme proposé et le nom considéré. Ils attribuent à chaque synonyme une note entre 1 et 5 selon que le fait de remplacer *sec* par ce synonyme dans le contexte du nom considéré provoque un changement de sens important ou pas. On attribue ensuite un taux de réussite permettant d'évaluer la performance de Visusyn. On considère que 1 et 2 sont des notes proches puisqu'elles signifient que l'adjectif considéré peut être utilisé comme synonyme, plus ou moins adéquat, de *sec* dans le contexte du nom considéré. De même 3 et 4 sont liées et signifient que l'adjectif est rejeté plus ou moins vivement. En revanche 3 et 2 sont faiblement liées par la notion de changement de sens. Le tableau ci dessous récapitule les règles d'attribution du taux de réussite.

Note sujet	1				2				3				4			
Note Visusyn	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
% de réussite	100	80	0	0	80	100	20	0	0	20	100	80	0	0	80	100

Une partie des résultats est présentée ci dessous. On a indiqué pour chaque adjectif l'arrondi de la moyenne des notes données par les sujets :

nom	**adjectif**	**note Visusyn**	**note sujets**	**réussite (%)**		**nom**	**adjectif**	**note visusyn**	**note sujets**	**réussite (%)**
coup	brusque	1	1	100		**terre**	brusque	4	4	100
	décharné	4	4	100			décharné	3	3	100
	desséché	4	4	100			desséché	1	1	100
	stérile	4	4	100			stérile	2	2	100
	glacial	4	4	100			glacial	3	4	80
			moyenne	**100**					**moyenne**	**96**

nom	adjectif	note Visusyn	note sujets	réussite (%)		nom	adjectif	note visusyn	note sujets	réussite (%)
fleur	brusque	4	4	100		**boue**	brusque	4	4	100
	décharné	4	3	80			décharné	4	4	100
	desséché	3	1	0			desséché	4	1	0
	stérile	2	4	0			stérile	4	4	100
	glacial	4	4	100			glacial	1	4	0
			moyenne	**56**					**moyenne**	**60**
manière	brusque	1	2	80		**lit**	brusque	3	4	80
	décharné	4	4	100			décharné	4	4	100
	desséché	4	4	100			desséché	4	1	0
	stérile	4	4	100			stérile	4	4	100
	glacial	4	2	0			glacial	3	4	80
			moyenne	**76**					**moyenne**	**72**

Le taux de réussite globale est de 79 %. Pour presque 63 % des cas considérés, le taux de réussite est supérieur à 70 %, ce qui veut dire que pour ces noms le calcul automatique du sens aboutit à la sélection d'un sens de *sec* valable dans le contexte du nom considéré.

Ainsi nous pouvons calculer correctement le sens de *mouvement sec (**brusque**), coup sec (**brusque**), main sèche (**desséchée**), corps sec (**décharné**), terre sèche (**desséchée / stérile**), ton sec (**brusque**), éclair sec (**brusque**), torrent sec (**desséché**), cou sec (**décharné / desséché**), manières sèches (**brusques**), sol sec (**desséché**), arbre sec (**décharné / desséché**).*

Dans la majorité des cas, nous pouvons déterminer s'il s'agit d'un synonyme parfait ou approximatif (attribution correcte de la note 1 ou 2) : *mouvement sec* et *mouvement brusque*, *coup sec* et *coup brusque*, *corps sec* et *corps décharné*, *terre sèche* et *terre desséchée*, *éclair sec* et *éclair brusque*, *cou sec* et *cou décharné* sont parfaitement synonymes. En revanche pour *main sèche* et *main desséchée*, *terre sèche* et *terre stérile*, *cou sec* et *cou desséché*, le remplacement de *sec* par son synonyme provoquera un léger changement de sens. Tout cela notre calcul le prédit parfaitement. Pour les deux noms *coup* et *éclair* la réussite est même parfaite (100 %), c'est à dire que non seulement le calcul sélectionne le bon sens de *sec* en présence du nom, mais avec la même précision que les sujets (note 1), et de plus il rejette aussi nettement les sens qui ne conviennent pas. Pour les noms *terre*, *arbre et cou* on a obtenu les deux sens possibles sélectionnés par les sujets (*terre desséchée / stérile*, *arbre ou cou décharné / desséché*). Cependant certains cas d'indétermination nous échappent (*ton brusque / glacial*, *manière brusques / glaciales*).

L'analyse détaillée des erreurs nous a permis de dégager plusieurs voies de travail en vue d'améliorer notre système. Nous travaillons actuellement à agrandir le corpus étudié. Le fait qu'on ait eu accès à un nombre limité de textes, en majorité très littéraires, explique que certains sens d'un usage plus quotidien échappent à notre calcul. C'est le cas de *manières glaciales* par exemple. Parallèlement à l'élargissement du corpus, il nous faut perfectionner notre mode de représentation et de calcul. Nous devons donner plus de poids aux cliques centrales, qui, bien que possédant peu de synonymes, représentent des sens très importants du mot vedette et qui doivent donc prendre plus de poids dans les calculs. C'est pour cette raison que nous ne réussissons pas à calculer le sens *fleur desséchée* pourtant attesté par les sujets. Il faudra étudier en détails quelle est la façon la plus efficace de rééquilibrer notre représentation. Un autre biais dans le calcul est du au fait que la présence de cliques ayant un fort degré d'affinité avec un nom donné peut être contrebalancée par le voisinage de cliques ayant un

degré d'affinité très faible. Une piste possible pour la résolution de ce problème est d'augmenter le nombre de dimensions de l'espace de représentation. Il semblerait intéressant aussi d'affiner notre construction de l'espace sémantique en utilisant d'autres relations paradigmatiques comme l'antonymie, voire en croisant divers indices de proximité entre synonymes comme l'ont fait Inkpen et Hirst (2003).

Enfin, le problème le plus important que nous ayons rencontré vient de ce que deux synonymes de *sec* peuvent se trouver en cooccurrence avec un même nom sans pour autant être dans ce cas synonymes entre eux. Nous avons rencontré le problème une première fois dans le cas de *lit*. L'analyse de ce cas nous a conduit à chercher quelles étaient les cliques ayant un degré d'affinité avec *lit* supérieur à 90 %. Ce sont toutes des cliques contenant un des adjectifs *dur*, *rude* ou *froid*. Or *lit froid*, *dur*, ou *rude* ne sont pas synonymes de *lit sec*. On se heurte ici au problème de la polysémie de *lit*, « meuble pour dormir » ou « cours de rivière ». Or associer *sec* à *lit* oblige, du moins dans le corpus étudié, à choisir le sens « lit de rivière » et par la même à exclure pour *sec* les sens *froid*, *dur* et donc *brusque*. La question qui se pose ici est : comment notre système va t'il opérer cette double sélection ? Sans doute d'autres éléments du contexte entreront-ils en ligne de compte dans le calcul, mais au bout du compte il lui faudra bien décider qu'un « lit de rivière » ne peut être qu'*asséché*. D'où tiendra t'il ce savoir ? De quel type d'apprentissage ?

On trouve un problème un peu similaire avec *boue*. Visusyn considère que *boue sèche* et *boue glaciale* sont synonymes, du fait que *boue* est à la fois très compatible avec *glacial* et avec *sec*. Notons que l'on rencontre le même phénomène avec *temps* : *temps sec* n'est pas synonyme de *temps glacial*. Ici la polysémie du nom n'entre pas en jeu. Le problème se pose d'abord parce que *sec* peut prendre des sens dépendant de domaines différents : les uns sont des sens physiques (*un arbre sec, du sable sec*), les autres psychologiques (*un cœur sec*), ensuite parce que parmi ses synonymes, certains comme *froid*, *glacé* et *glacial* peuvent aussi déployer leur sens dans les deux domaines (*une eau froide / un abord froid, une boisson glacée / un accueil glacé, un vent glacial / un sourire glacial*), enfin parce que *sec*, *froid*, *glacial*, *glacé* ne sont synonymes que dans leurs sens psychologiques mais que certains noms, aussi utilisés avec *sec*, peuvent se trouver en cooccurrence avec eux dans un sens physique qui échappe à la synonymie avec *sec*. Notons d'autre part que si ce phénomène a une incidence notable sur nos calculs, c'est aussi parce que ces adjectifs partagent de nombreuses cliques. En effet notre méthode de calcul est robuste et lorsqu'un tel phénomène ne concerne qu'un nombre limité de cliques, il n'interfère pas dans les calculs.

C'est ce genre de problèmes qui nous ont convaincus qu'une étude théorique mathématique et informatique plus approfondie de la structure du graphe de synonymie était nécessaire pour progresser. Pour prendre en compte des interactions aussi subtiles que celle de *sec*, *glacial* et *froid*, il faudrait pouvoir obtenir une visualisation plus globale du graphe de synonymie. Notre méthode de représentation à partir des cliques, dans laquelle les unités lexicales occupent des régions plus ou moins grandes suivant leur polysémie est assez efficace mais essentiellement locale. On ne peut visualiser le graphe au voisinage d'un sommet. Nous cherchons actuellement, en collaboration avec Bruno Gaume (Irit Toulouse) à définir des méthodes permettant de visualiser des parties plus importantes du graphe. En utilisant des structures plus « lâches » que les cliques, les « gangs », on peut obtenir des cartes du graphe à différentes échelles, de la plus locale à la plus globale. Nous voudrions construire ainsi une sorte d'atlas sémantique du français.

5. Conclusions et perspectives

Les résultats de l'expérience sont très encourageants. Nous avons un taux de réussite de 79 %. Nous pouvons raisonnablement considérer que, même sans résoudre immédiatement tous les problèmes théoriques, il va pouvoir atteindre les 90 %. Nos principales voies de travail sont actuellement l'agrandissement du corpus et le réajustement des paramètres de calcul, notamment à partir de résultats d'expériences psycholinguistiques menées à plus grandes échelle par des experts. En outre nous comptons beaucoup sur le travail de Jacquet (2003) sur la désambiguïsation des verbes et l'influence des constructions verbales dans la construction du sens. Nous pouvons donc espérer proposer rapidement une méthode automatique et complète de calcul du sens sur des grands corpus. Enfin les problèmes rencontrés avec *lit* et *boue* nous pousse à approfondir notre réflexion théorique. Dans le cas de *lit*, il s'agit de trouver comment la polysémie du nom régissant entre en jeu. C'est tout l'aspect dynamique du calcul du sens qui apparaît ici, *sec* et *lit* s'influencent mutuellement. Ce qu'il s'agit de développer maintenant, c'est un autre outil informatique à base de réseaux connexionnistes récurrents, qui calcule de manière automatique les contraintes réciproques des unités polysémiques apparaissant dans un même énoncé. D'autre part le travail entrepris sur les grands graphes devrait nous mener plus loin que l'avancée de nos travaux en modélisation de la polysémie. Les graphes sont de plus en plus utilisés en sémantique, notamment dans les études portant sur la connaissance lexicale. Les sommets de ces graphes représentent les mots d'une langue. Il existe plusieurs types de réseaux selon la relation sémantique utilisée pour définir les arcs du graphe. Celle ci peut être de type syntagmatique ou de cooccurrence : on construit un arc entre deux mots si on les trouve au voisinage d'un mot cible (Veronis, 2003). Elle peut être de type paradigmatique comme c'est le cas dans notre graphe de synonymie. Il peut s'agir d'une relation plus générale de proximité sémantique prenant en compte à la fois l'axe paradigmatique et l'axe syntagmatique (Gaume *et al.*, 2002). On peut enfin imaginer de relier des mots sur des critères distributionnels, suivant les contextes qu'ils partagent, comme le fait Bourigault (2002). Ces graphes, tout comme la plupart des grands graphes de terrain (graphes sémantiques, réseaux géographiques, électriques, Internet…) partagent une topologie bien particulière (peu denses présentant une structuration locale riche et une distance moyenne très petite sur l'ensemble du graphe ainsi qu'une structure hiérarchique) et on les appelle graphes de type « small world ».Les outils que nous allons mettre en place pourront donc être utilisés directement dans d'autres domaines des sciences sociales.

L'idée qui sous tend nos travaux est que ces graphes recèlent dans leur structure une information très riche mais difficile à utiliser directement. Notre méthode de géométrisation devrait permettre une meilleure compréhension des phénomènes sous jacents. On peut ainsi espérer en savoir plus sur la structure sémantique du lexique d'une langue et mettre au point des méthodes de navigation dans le lexique.

Références

Audibert. L. (2003). Étude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences. In *Actes de TALN 2003* : 35-44.

Bourigault D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de TALN 2002.*

Edmonds P. et Hirst G. (2002). Nearsynonymy and lexical choice. *Computational Linguistics*, vol. (28/2) :105-144.

Gaume B., Duvignau K., Gasquet O. et Gineste M.-D. (2002). Forms of Meaning, Meanings of Forms. *Journal of Experiment and Theoretical Artificial Intelligence*, vol. (14/1) : 61-74.

Guthrie J.A., Guthrie L., Wilks Y. et Aidinejad H. (19991). Subject-dependent co-occurrence and word sense disambiguation. In Morristown NJ (Ed.), *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* : 146-152.

Ide N. et Véronis J. (1998). Introduction to the special issue on word sense disambiguation : the state of the art. *Computational linguistic*, vol. (24/1) : 1-40.

Jacquet G. (2003). Polysémie verbale et construction syntaxique : étude sur le verbe *jouer*. In *Actes TALN 2003* : 469-479.

Kilgariff A. (1994). The myth of completeness and some problems with consistency (the role of frequency in deciding what goes in the dictionary. In *Proceedings of the 6th International Congress on Lexicography, EURALEX'94* : 101-106.

Inkpen D.Z. et Hirst G. (2003). Automatic sense disambiguation of the near-synonyms in a dictionary entry. In *Proceedings of the 4th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*.

Lesk M. (1986). Automated Sense Disambiguation : How to Tell Pine Cone from an Ice Cream Cone. In *Proceedings of the 19986 SIGDOC Conférence*. Association for Computing Machinery *:* 24-26.

Ploux S. et Victorri B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes. *TAL*, vol. (39/1) :161-182.

Ravin Y. et Leacock C. (2000). *Polysemy : Theoretical and Computational Approaches*. Oxford University Press**.**

Reymond D. (2001). Dictionnaires distributionnels et étiquetage lexical de corpus. In *Actes de RECITAL 2001 :* 24-33.

Véronis J. et Ide N. (1990). Word Sense Disambiguation with Very Large Neural Network Extracted from Machine Readable Dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics, COLING'90*, vol. (2) : 389-394.

Véronis J. (2003). Cartographie lexicale pour la recherche d'information. In *Actes de TALN 2003 :* 265-275.

Victorri B. et Fuchs C. (1996). *La polysémie, construction dynamique du sens*. Hermès.

Victorri B. (2002). Espaces sémantiques et représentation du sens. *Textualités et nouvelles technologies*, éc/artS, 3.

Walker D.E. (1987). Knowledge resource tools for accessing large text files. In Nirenburg S. (Ed.), *Machine Translation*. Cambridge University Press.

Wilks Y.A. et Fass D. (1990). *Preference semantics : A family history*. Report MCCS-90-194, Computing Research Laboratory, New Mexico State University, Las Cruces.

Wilks Y.A., Fass D., Guo C.-M., McDonal J.E., Plate T. et Slator B.M. (1993). Machine tractable dictionary tools. In Pustejovsky J. (Ed.), *Semantics and the Lexicon*. Kluwer.

Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource

Jacques Vergne

GREYC – Université de Caen – BP 5186 – 14032 Caen cedex
www.info.unicaen.fr/~jvergne

Abstract

Our present research is in the field of exploring NLP methods using no other resource than the text to analyse itself. This drives us to analysis methods which use very general linguistic properties, as for instance differences of length and frequencies of words. To illustrate our approach, we present in this paper a method of local computation for discovering function words from raw corpora. This method can be used for extracting term candidates or indexing raw texts of unidentified alphabetic natural languages.

Résumé

Nous nous plaçons dans une perspective de traitements linguistiques sans autre ressource que le texte à analyser. Ceci nous conduit à des méthodes d'analyse exploitant des propriétés très générales des langues, comme par exemple les différences de longueur et de fréquence des mots. Pour illustrer notre démarche, nous présentons dans cet article une méthode de découverte des mots vides par un calcul local. Cette méthode peut s'appliquer à l'extraction de candidats termes ou à l'indexation de textes bruts de langues alphabétiques non identifiées.

Mots-clés : traitements multilingues, découverte des mots vides, multilingual NLP, natural language learning, grammar induction, function words discovery.

1. Introduction

Ce travail se situe dans le cadre d'une exploration dans la direction de traitements linguistiques sans aucune autre ressource que le texte à analyser lui-même. Dans ce cas, l'absence de ressources permet d'envisager qu'un même analyseur puisse traiter des textes de langues différentes, sans identification de la langue. Bien évidemment, un tel analyseur doit exploiter des propriétés linguistiques très générales, et non pas des propriétés locales à une langue, telles qu'un lexique monolingue. Cette exploration sur les traitements s'accompagne donc aussi d'une exploration de propriétés linguistiques communes à un groupe de langues.

Dans cet article, nous présentons une méthode de découverte des mots vides, méthode pouvant par exemple prendre place dans le cadre d'une tâche d'extraction de candidats termes (Vergne, 2003), ou d'indexation automatique. De telles tâches nécessitent de repérer les expressions nominales fréquentes. Habituellement, on pratique un étiquetage des mots, ou une analyse morphosyntaxique (Bourigault, 2002), ou bien on dispose d'un « anti-dictionnaire » des mots vides pour pouvoir sélectionner les segments fréquents qui ne sont pas des mots vides (Salem, 1987 ; Salton *et al.*, 1993 ; Ahonen-Myka, 1999). Dans les deux cas, le corpus est supposé monolingue et la langue identifiée. Nous présentons ici une méthode sans ressource linguistique, fonctionnant sur corpus écrits bruts monolingues ou multilingues, de langues alphabétiques non identifiées.

La tâche que nous nous proposons est de prendre un tel corpus, et de découvrir les mots vides **présents dans ce corpus**. La tâche plus générale de découverte des structures syntaxiques à partir de corpus bruts a déjà été explorée (et partiellement résolue) par Hervé Déjean (Déjean, 1998), et ce travail est actuellement poursuivi dans le courant de la « Grammatical Induction of languages » (ou Grammar Induction, Grammatical Inference, ou Grammar Inference[1]).

2. Mots vides, mots grammaticaux, stop-list et stopwords

Nous avons choisi les termes : « mot vide » – « mot plein », synonymes de « mot grammatical » – « mot lexical » et de « function word » – « content word », à la suite de Lucien Tesnière (1969 : 53) et de Fathi Debili (1982).

Dans la tradition de l'informatique documentaire, la définition est plus extensive : un « mot vide » est un mot qui ne doit pas être indexé, qu'il soit mot grammatical ou mot lexical non discriminant (thème commun à une base documentaire). Les mots vides sont alors souvent regroupés dans un « anti-dictionnaire » ou une « stop-list » ou une liste de « stopwords ». Il est généralement admis que ces mots très fréquents (environ la moitié des occurrences d'un texte) ne sont pas à indexer, car ils ne sont pas informatifs, et ils augmentent énormément la taille de l'index si l'on réalise une indexation fulltext, ce qui est le cas courant. Des auteurs au contraire prônent l'indexation des mots vides, comme pouvant être informatifs, les prépositions par exemple (Riloff, 1995).

Dans la littérature, nous n'avons pas trouvé de méthode de découverte des mots vides dans les corpus bruts (hormis celle d'Hervé Déjean, fondée sur la découverte des morphèmes et leur étude statistique positionnelle). Par contre, on trouve des travaux où une liste de mots vides placée en entrée constitue l'amorce d'un traitement :

- Pour Wilbur et Sirotkin, dans leur article « *The automatic identification of stop words* » (Wilbur et Sirotkin, 1992), un « stop word » est tout mot non informatif dans une collection de documents. Ils amorcent leur algorithme par une « *standard stopword list* ». Yiming Yang (1995) décrit leur méthode : « *In contrast to using generic stop words, Wilbur and Sirotkin developed a novel stopword identification method which allows a far more aggressive removal of words from documents without losing retrieval accuracy.* »
- Tin Kam Ho, dans son article « *Fast Identification of Stop Words for Font Learning and Keyword Spotting* » (Ho, 1999), a pour premier objectif d'améliorer l'apprentissage des polices dans un système d'OCR, en concentrant le début de cet apprentissage sur les mots vides, et en cherchant à apparier les graphies des mots vides placés en entrée, et les images des mots courts délimités dans l'image du document. Remarquons que l'auteur utilise la longueur des images de mots (« *word width* ») comme critère discriminant vide - plein, mais sans citer Zipf.
- Andrew Roberts, dans « *Automatic Acquisition of Word Classification using Distributional Analysis of Content Words with Respect to Function Words* » (Roberts, 2002) propose « *a method which can automatically infer word classification* » à partir d'une liste de mots vides placée en entrée. Il mentionne qu'il aurait pu les obtenir automatiquement, et à ce propos, il cite la méthode d'Eliott (Elliott *et al.*, 2000), du même laboratoire : « *Elliott has found that function words can be obtained by combining relatively small samples from at least 3 sources.* » L'objectif (surprenant) des auteurs est de détecter un matériau linguistique dans des signaux venant de l'espace, et, pour eux, cette tâche inclut la découverte de

[1] Voir le «Grammatical Induction Community website» : `http://eurise.univ-st-etienne.fr/gi/`

mots vides ; pour ce faire, ils proposent d'utiliser la « *least inter-text variation* » des mots vides, à partir de l'hypothèse (non explorée, ni validée) que les mots vides sont plus stables que les mots pleins dans des corpus thématiquement différents.

3. Propriétés linguistiques

Des propriétés linguistiques très générales sont utilisées pour catégoriser les mots vides ou pleins : des propriétés de longueur et de fréquence des mots, et de différences de ces critères entre deux mots contigus :

- Une observation fondamentale de Zipf est que les mots pleins sont rares et longs et que les mots vides sont **plus fréquents et plus courts** : ce qui est d'usage fréquent est court : c'est la loi de l'économie d'effort dans l'usage d'un code, caractérisée par Zipf (1949), et observable aussi dans les langages de programmation (remarquons que la « loi de Zipf », toujours très présente dans la littérature, est une loi sur les effectifs des mots, et que les propriétés statistiques des longueurs des mots sont plus rarement invoquées).
- En application de la propriété énoncée par Saussure : « *dans la langue, il n'y a que des différences* » (Saussure, 1922, éd. 1974 : 166), nous allons fonder nos calculs sur les **différences locales** de longueur et d'effectif de mots contigus.

Une autre propriété linguistique très générale nous conduit à rechercher une méthode qui n'utilise pas de « stoplist » : une même graphie peut coder deux mots **homographes** différents dans des contextes différents, et ces deux mots peuvent devoir être catégorisés l'un vide, l'autre plein. De telles graphies sont fréquentes, par exemple en français : *car*, *or*, *la*, *son*, *une* (dans notre corpus de sites de presse), *pendant*, *avions*, ou *like* en anglais. Si l'on se fie à la valeur absolue des graphies d'une stoplist, on n'indexe pas ces mots, d'où un silence systématique. On doit au contraire se fier au contexte en faisant un **calcul local**, pouvant donner pour une même graphie des résultats différents selon le contexte.

4. Processus de découverte des mots vides par calcul local

Une étape préliminaire compte le nombre d'occurrences dans le corpus, de chaque graphie du corpus. Puis, le corpus est découpé en segments de corpus ne contenant pas de ponctuation.

Chaque segment est ainsi traité :

- il est d'abord segmenté en mots ;
- puis on recherche un **pavage** du segment avec les motifs suivants : `PvvvP`, `PvvP`, `PvP`, `vPv`, `vPPv`, `vPPPƒ`, `vvPƒ`, `vPPƒ`, `∂vvvP`, `∂vvP`, `∂vP`, où `v` = vide, `P` = plein, `ƒ` = en fin de segment, `∂` = en début de segment.

Pour avoir une complexité linéaire, nous avons choisi un pavage déterministe : plusieurs pavages seraient possibles, mais le premier motif validé est accepté. Les motifs sont essayés dans l'ordre ci-dessus. Les motifs peuvent se recouvrir sur 1 ou 2 mots. Si un motif s'applique, les mots qu'il contient prennent la catégorie correspondante du motif. Si aucun motif ne s'applique, les mots qu'il contient restent indéterminés.

Un motif est validé par la règle suivante : dans le motif, l'ensemble des mots vides doit être différent de l'ensemble des mots pleins ; c'est-à-dire que pour chaque critère (les longueurs en nombre de lettres et les effectifs), la moyenne géométrique des valeurs minimale et maximale du critère sépare les mots vides des mots pleins du motif.

Exemple de validation d'un motif sur le début d'un segment :

```
0          2-189 le
1          6-41  nombre      pour chaque mot : longueur-effectif
2          3-384 des
```

- le motif `PvvP` est invalidé car on n'a pas la différence plein - vide sur les mots `0 1`, car on a : 2<6 (sur les longueurs)
- le motif `PvP` est invalidé pour la même raison
- validation du motif `vPv` :

. on a une coupe entre pleins et vides selon les effectifs :
moyenne géométrique (41 ; 384) = **125,46**
`41 (P) < 125.46 < 189 (v) 384 (v)`

. on a une coupe entre pleins et vides selon les longueurs :
moyenne géométrique (2 ; 6) = **3,46**
`6 (P) > 3.46 > 3 (v) 2 (v)`

le motif `vPv` est donc validé, et les 3 occurrences sont catégorisées vide - Plein - vide :

```
0     v     2-189 le
1     P     6-41  nombre
2     v     3-384 des
```

Voici un exemple de résultat pour un segment :

```
---PP-P     P est imposé par la graphie (si longueur > 6)
vPv         motifs
.PvP            validés
..vPPv              au cours du pavage
....PvP
vPvPPvP     résultat du pavage => catégorie locale pour chaque occurrence de mot
0     v     2-189 le
1     P     6-41  nombre      pour chaque mot : longueur-effectif
2     v     3-384 des
3     P     7-10  lycéens
4     P     9-5   préparant
5     v     2-75  un
6     P     12-21 baccalauréat
```

La complexité de l'algorithme est linéaire en temps selon le nombre de mots du corpus.

5. Résultats et évaluation

Voici des exemples de résultats sur trois corpus monolingues anglais, français et allemand de tailles analogues (en Ko) et de même genre (textes de pages acquises automatiquement par crawling de sites de presse), suivis d'une évaluation comparative sur les trois corpus. Nous donnons quelques exemples de mots homographes de catégories différentes.

5.1. *Corpus monolingue anglais*

Dans ces deux segments, « `like` » est mot vide ou mot plein :

```
0     v     4-27  like        ici, « like » est un mot vide
1     P     6-1   bamboo
2     P     6-1   shoots
```

```
3     v     5-11  after
4     v     1-252 a
5     P     6-1   spring
6     P     4-1   rain

0     v     3-33  But
1     v     2-37  we
2     P     4-27  like        ici, « like » est un mot plein
3     v     2-289 to
4     P     3-4   buy
5     -     5-16  those       occurrence indéterminée
6     P     10-8  businesses
7     v     2-249 in
8     v     1-252 a
9     P     10-1  contrarian
10    P     7-1   fashion
```

5.2. *Corpus monolingue français*

Dans ces deux segments, « une » est mot plein ou mot vide :

```
0     P     10-12 ACTUALITES
1     v     1-201 à
2     v     2-323 la
3     P     3-119 une         ici, « une » est un mot plein

0     v     3-273 les
1     P     8-3   Français
2     P     11-1  rencontrent
3     v     6-15  samedi      bruit sur les mots vides
4     v     3-119 une         ici, « une » est un mot vide
5     P     6-1   équipe
6     v     2-140 du
7     P     6-3   Canada
8     v     1-117 a
9     P     6-1   priori
10    v     1-201 à
11    v     2-16  sa
12    P     6-2   portée
13    v     2-196 et
14    P     8-1   composée
15    v     2-143 en
16    P     8-12  majorité
17    v     2-531 de
18    P     7-3   joueurs
19    P     5-3   ayant       silence sur les mots vides
20    v     4-22  fait        bruit sur les mots vides
21    v     2-206 le
22    P     5-2   choix
```

5.3. *Corpus monolingue allemand*

```
0     v     3-144 Die
1     P     6-2   zweite
```

```
2     P     8-1   Vorrunde
3     v     3-234 der
4     P     19-1  Ausscheidungsrennen
5     v     3-20  zum
6     P     12-1  diesjährigen
7     P     7-3   America
8     v     1-13  s
9     P     3-4   Cup
10    v     3-39  ist
11    v     2-66  zu
12    P     4-7   Ende
13    P     8-3   gegangen
```

5.4. *Évaluation comparative sur les 3 corpus anglais, français, allemand*

Définissons le taux global de détermination : 1 - (nombre d'occurrences de mots non catégorisées / nombre total d'occurrences).

L'évaluation est faite du point de vue de l'extraction des mots vides. Pour ce faire, nous allons définir une occurrence locale de mot vide de manière opératoire : un mot vide n'est pas un mot plein ; un mot plein est un nom, un adjectif non numéral, un verbe non auxiliaire et non modal, ou un adverbe.

Comme une même graphie peut être catégorisée différemment dans des contextes différents, seule l'évaluation sur les nombres d'occurrences a un sens (et non pas l'évaluation sur les nombres de graphies différentes) :

- précision = 1 - bruit = 1 - (nombre d'occurrences de mot plein catégorisées vides / nombre d'occurrences de mot vide extraites)
- rappel = 1 - silence = 1 - (nombre d'occurrences de mot vide non catégorisées vides / nombre total d'occurrences de mot vide).

langue	**site de presse** (23/11/02)	**taille**	**détermi -nation**	**précision**	**rappel**	**longueurs**	
						moyenne	écart-type
anglais	The International Herald Tribune	76 Ko, 12 501 mots	90,8%	92,6%	78,2%	4,51	2,73
français	Le Monde	82 Ko, 12 348 mots	94,2%	96,3%	85,8%	4,96	3,14
allemand	Der Spiegel	81 Ko, 9 897 mots	92,8%	90,8%	88,7%	5,99	3,71

Dans le corpus anglais, on observe que les mots sont plus petits et leurs longueurs peu dispersées, alors que dans le corpus allemand, les mots sont plus longs et les longueurs plus dispersées (mots composés). On peut observer en outre une corrélation entre rappels et écart-types des longueurs. Hypothèse : le rappel est meilleur si les longueurs sont plus dispersées, ce qui permet un meilleur contraste dans la validation d'un motif du pavage. Ceci pourrait expliquer pourquoi le rappel est meilleur en allemand.

6. Discussion

Interrogeons-nous sur les deux critères utilisés : l'effectif et la longueur. Ce sont des caractéristiques globales de la graphie, de même valeur pour toutes les occurrences de cette graphie.

C'est seulement la séquence locale de quelques occurrences de graphies qui permet d'affecter une valeur locale, vide ou plein, à une occurrence, par un calcul local sur des valeurs globales. Au sujet de la longueur d'un mot, Zipf (1935) choisit la syllabe ou le phonème comme unité de la métrique, sans motiver ce choix ; en ce qui nous concerne, nous devons choisir une unité indépendante des langues, donc la lettre, c'est-à-dire directement la forme écrite de la graphie, sans le calcul (dépendant de la langue) qu'aurait nécessité la syllabation ou la phonétisation. Nous avons fait des tests en choisissant la syllabe, pour le français, avec des résultats analogues. Nous avons aussi fait des tests en ne prenant que les effectifs ou que les longueurs, avec des résultats corrects mais inférieurs, marqués par plus de bruit, mais moins de silence (sur les mots vides).

Pour qu'il y ait une détection (locale) d'un mot vide entre deux mots pleins ou inversement, il faut que les différences soient ensemble **suffisamment contrastées**. Cette condition n'est pas satisfaite en moyenne pour environ 10% des occurrences d'un mot vide (c'est la cause principale du silence), quand un mot vide est long et/ou rare, ou quand un mot plein voisin est court et/ou fréquent. Dans une phase ultérieure d'analyse (non décrite ici, mais en cours de réalisation), on peut appliquer localement ce qui a été découvert ailleurs. C'est tout le problème de l'application globale de certaines déductions locales ; on doit satisfaire à deux contraintes contradictoires : une même forme doit pouvoir obtenir plusieurs catégories, et appartenir à plusieurs paradigmes de mots vides (surtout si le corpus est multilingue), ce qui interdit la généralisation totale, mais il est nécessaire de pratiquer une généralisation partielle pour diminuer le silence des mots vides (ce point est à l'étude).

Sur quelles langues (alphabétiques) cette méthode donne-t-elle de bons résultats ? Les tests conduits sur d'autres langues nous conduisent à répondre : sur celles dont les « mots » vides sont des mots isolés (italien, espagnol, suédois par exemple) et non pas des morphèmes accolés à des mots pleins (langues agglutinantes telles que turc, finnois par exemple). Dans ce dernier cas, un prétraitement d'isolation des morphèmes est envisageable (comme l'a montré Hervé Déjean, 1998).

7. Conclusion et perspectives

Nous avons présenté une méthode de découverte des mots vides dans des corpus bruts de langues alphabétiques non identifiées, par calcul local. Une telle méthode constitue un exemple de traitement sans autre ressource que le texte analysé. De tels traitements, que l'on pourrait appeler « alingues », doivent exploiter des propriétés très générales des langues, propriétés qu'il s'agit d'expliciter et valider. La bonne détection des mots vides est un indice de la généralité des propriétés linguistiques exploitées. La direction des traitements linguistiques sans ressources est prometteuse.

Nos travaux actuels portent sur la phase d'analyse dans les directions suivantes : généraliser partiellement les déductions locales, catégoriser les mots vides, réaliser le chunking, et distinguer les chunks nominaux des chunks verbaux. Le chunking « alingue » sans ressources est maintenant envisageable. Nous devons aussi augmenter le nombre de langues, ce qui nécessite un locuteur de chaque langue pour l'évaluation.

Références

Ahonen-Myka H. (2002). Discovery of frequent word sequences in text. In *The ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining*. www.cs.helsinki.fi/u/hahonen/ahonenmyka_patws02.ps

Bourigault D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de TALN 2002* : 75-84.
www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TALN02-Bourigault.doc

Debili F. (1982). *Analyse syntaxico-sémantique fondée sur une acquisition automatique des relations lexicales-sémantiques*. Thèse de doctorat d'état en sciences informatiques Université de Paris XI.

Déjean H. (1998). *Concepts et algorithmes pour la découverte des structures formelles des langues*. Thèse en informatique, Université de Caen.

Elliott J., Atwell E. et Whyte B. (2000). Language identification in unknown signals In. *Proceedings of CoLing'2000, 18th International Conference on Computational Linguistics* : 1021-1026.

Grammatical Induction Community website : http://eurise.univ-st-etienne.fr/gi/

Ho T. K. (1999). Fast Identification of Stop Words for Font Learning and Keyword Spotting. In *Proceedings of the 5th Int'l Conference on Document Analysis and Recognition*.

Riloff E. (1995). Little Words Can Make a Big Difference for Text Classification. In *Proceedings of {SIGIR}-95, 18th {ACM} International Conference on Research and Development in Information Retrieval* : 130-136.

Roberts A. (2002). *Automatic Acquisition of Word Classification using Distributional Analysis of Content Words with Respect to Function Words*. School of Computing, University of Leeds.

Salem A. (1987). *Pratique des segments répétés*. Klincksieck.

Salton G. and Allan James (1993). Selective Text Utilization and Text Traversal. In *UK Conference on Hypertext* : 131-144.

Saussure F. de (1922, éd. 1974). *Cours de Linguistique Générale*. Payot.

Tesnière L. (1982). *Éléments de syntaxe structurale*. Klincksieck. (1re édition: 1959)

Vergne J. (2003). Un outil d'extraction terminologique endogène et multilingue.In *Actes de TALN 2003*, tome (2) : 139-148.
www.info.unicaen.fr/~jvergne/TALN2003/JVergne-TAL2003multV23.pdf

Wilbur JW et Sirotkin K. (1992). The automatic identification of stop words. *Journal of Information Science*, 18(1) : 45--55.

Yang Y. (1995). Noise Reduction in a Statistical Approach to Text Categorization. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval*.

Zipf G.K. (1935). *The Psychobiology of Language, an Introduction to Dynamic Philology*. Houghton Mifflin.

Zipf G.K. (1949). *Human Behavior and the Principle of Least Effort*. Harper. (Réédition 1966).

Analyse séquencée de la micro-distribution lexicale

Jean-Marie Viprey

Archives, Textes, Sciences des Textes
Université de Franche-Comté – 25030 Besançon Cedex – France
jean-marie.viprey@univ.fcomte.fr

Abstract

Once one has worked out macro- and micro-distribution of the lexical items in the observed corpus, one wishes to study the collocation behavior of specific items, concerning diachrony and/or contrast. The *sequenced* analysis of micro-distributions consists in collecting the occurrence contexts of a chosen item, distributing their elements into the relevant sections of the corpus. The matrix made by this way can pass FCA, which gives a first synthetic view over evolution and/or contrasts inside signification. This method is to be integrated into expert hypertext, in order to increase our knowledge of the corpus, to rationalize critical readings by selecting relevant extracts, and to build ans supervise hypotheses. Applications to the *Monde Diplomatique* (1980-2000) and to the authorship dispute about Corneille and Molière.

Résumé

Une fois établies la macro-distribution et la micro-distribution des items lexicaux dans le corpus à l'étude, on souhaitera observer le comportement co-occurrentiel d'items ou de groupes d'items particuliers, dans une optique diachronique et/ou contrastive. L'analyse *séquencée* des micro-distributions consiste à relever les contextes d'occurrence du pivot et à en distribuer les éléments dans les divisions pertinentes du corpus. La matrice ainsi constituée est soumise à l'AFC, qui donne une première vue synthétique de l'évolution et/ou des contrastes dans la signification. Cette méthode s'intègre à l'hypertexte expert afin de favoriser la connaissance du corpus, de rationaliser les retours au texte par extraits pertinents, et de construire, affiner ou vérifier les hypothèses d'interprétation. Application au *Monde Diplomatique* (1980-2000) et à la controverse d'attribution sur Corneille et Molière.

Mots-clés : analyse multi-dimensionnelle, vocabulaire, diachronie, cooccurrence.

1. Introduction

Viprey (1997, 2000 et 2002a) montre que l'AFC appliquée à des matrices de cooccurrence permet d'asseoir une cartographie lexico-thématique pour l'hypertexte expert dans les bases et corpus littéraires. Viprey (2002b) en est une application aux *Fleurs du mal* de Baudelaire.

Le principe en est de substituer et/ou combiner, aux hypothèses exogènes, des visualisations endogènes sur la structure lexico-thématique afin d'accentuer les vertus exploratoires de l'environnement critique. Ces vues, dont les items sont sensibles à diverses modalités de clic souris, offrent le continuum des relations *isotropiques* (*trepein : pencher vers*), parentés des profils micro-distributionnels.

Par rapport à l'état de l'art du milieu des années 90, nous avons donc cherché à privilégier la micro-distribution, jusqu'alors masquée par la macro-distribution (distribution des items dans les sous-ensembles linéaires du corpus), condition *sine qua non* de l'extraction des structures fines qui SONT le *vocabulaire* comme structure de *textualité*.

On a pu légitimement reprocher à cette méthode de rejeter dans l'ombre les perspectives diachroniques et contrastives. Sans doute ce rejet a-t-il d'abord été voulu, afin de grossir le trait quant au changement de perspective recherché. Mais nous avons très tôt œuvré à la réintégration des critères de temps et de contraste, indispensables à toute entreprise d'analyse des discours. Il n'était pas moins nécessaire que cette réintégration se fasse dans le cadre de la méthode nouvelle et non par retour aux techniques antérieures. Aux JADT 2002, nous avons présenté un premier moyen dédié aux corpus à dimension diachronique : une animation présentant les transitions entre *n* AFC micro-distributionnelles calculées sur les tranches successives chevauchantes du corpus.

Des travaux plus récents, notamment sur des corpus de discours institutionnels fortement diachroniques, ont exigé la mise au point d'un autre outil dans le cadre de ®Astartex. Il s'agit d'une combinaison plus étroite et plus directement lisible de la micro- et de la macro- distributions, qui permet de visualiser les aspects les plus remarquables de l'évolution diachronique de la cooccurrence autour d'un pivot lexical.

Nous en présentons ici le principe et deux applications, l'une dans le domaine du discours politico-médiatique (exploration du corpus du *Monde Diplomatique* 1980-2000), l'autre dans la sphère littéraire (apport d'éléments de jugement dans le cadre du dossier de l'attribution de certaines pièces de Molière à Corneille).

2. Principes du repérage et application au *Monde Diplomatique*

Le corpus d'étude est celui du mensuel *Le Monde Diplomatique*, de 1980 à 2000 soit 21 années, 7000 articles pour 17 millions de mots environ. Il n'est pas étiqueté.

La distribution des items les plus récurrents par année d'édition est éloquente (fig. A1, en annexe[1]). Les positions respectives d'items lexicaux tels que (*travailleurs, gauche, état*) et de groupes d'années très bien délimités (80 à 83 par exemple), nous donnent des indications précieuses, tant sur la signification dont ces items lexicaux se chargent dans la configuration de ce corpus, que sur la signification globale du corpus : à partir de nos connaissances sur la place qu'occupe *Le Monde Diplomatique* dans le discours politique et social, que pouvons-nous supposer de l'évolution générale de ce discours ? Compte tenu de ce que nous savons sur l'évolution du discours politique et social de 1980 à 2000, comment pouvons-nous repenser la position qu'y occupe *Le Monde Diplomatique* ?

La micro-distribution de ces mêmes items (dans les limites de la phrase) est tout aussi intéressante (fig.A2, en annexe). On y voit clairement apparaître les dominantes lexico-thématiques les plus significatives (on n'a retenu ici que les 2 premiers facteurs sur plus de 200). On constate que la ventilation des items sur chacun des deux graphes est très différente ; elle a d'ailleurs une signification évidemment différente.

Nous constatons que *femmes* (qui est aussi un pôle du graphe macro-distributionnel) occupe une position polaire sur le graphe micro-distributionnel d'ensemble. Cette forme a un profil voisin de ceux de *social(e)(s), société, enfants, jeunes, vie, culture, travail*, etc. Elle appartient à cette *isotropie* à l'échelle de l'ensemble du *Monde diplomatique* 1980-2000. Mais comment en savoir plus sur la signification diachronique de cette forme ?

Nous nous proposons de relever les contextes des 3603 occurrences de *femmes*, dans les limites de la phrase toujours (et de 25 mots à gauche et à droite), et de distribuer les coccur-

[1] Les AFC de cette étude sont réalisées avec le programme écrit dans SciLab par Alain Lelu (LASELDI, Université de Franche-Comté).

rences dans une matrice diachronique (une colonne par année). Un seuil de cooccurrence est fixé (au moins égal à 33 % du nombre de partitions – années, dans la limite des 200 premiers cooccurrents).

L'AFC de cette matrice nous éclaire sur la tendance cooccurrentielle de *femmes*, donc sur un aspect important de l'évolution de sa signification, dans *Le Monde Diplomatique* (et dans la mesure du possible, dans le discours politique et social tel que ce mensuel y contribue). On y voit se dessiner deux zones polaires assez bien différenciées, correspondant l'une aux années 80 à 83, l'autre aux années 95 à 2000, les années intermédiaires se distribuant de manière plus confuse..

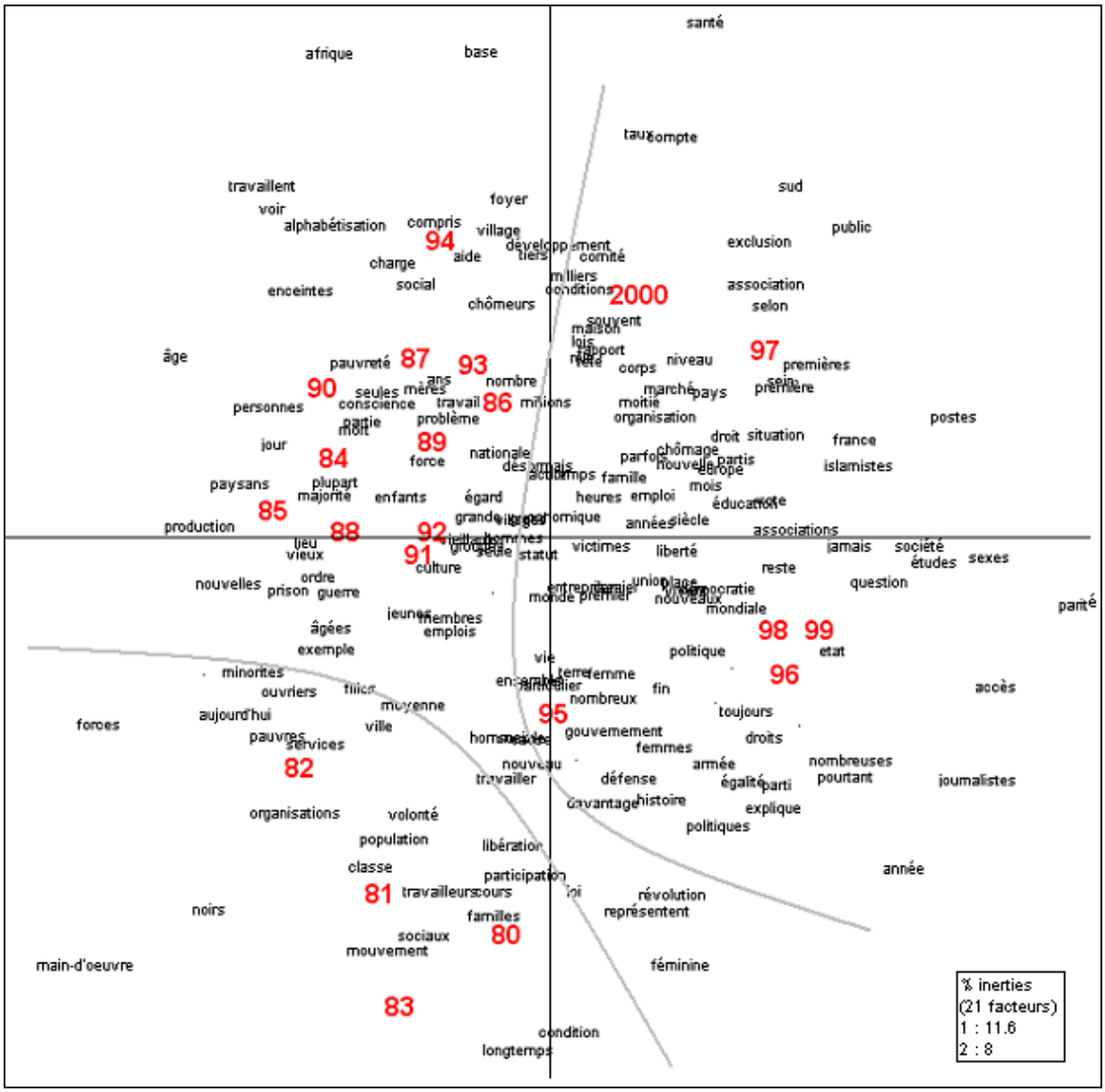


Figure 1. Graphe macro-distributionnel des cooccurrents de femmes*. 186 formes lexicales*

L'intérêt majeur de ces graphes étant de s'intégrer à un hypertexte expert, facilitant le retour immédiat et structuré aux contextes pertinents, nous ne pouvons en donner ici qu'une idée trop générale. Il est néanmoins patent que les cooccurrents caractéristiques des années 80-83 font référence à la sphère des rapports sociaux « classiques » telle qu'elle se définissait encore

alors, tandis que ceux de 95-2000 semblent se rapporter plus confusément aux questions de statut.

Si l'on détermine, par l'écart-réduit à l'équidistribution des cooccurrences, quels sont les plus forts cooccurrents de *femmes* dans les quatre périodes successives de cinq ans :

1980-84	cc	er	1985-90	cc	er	1991-95	cc	er	96-2000	cc	er
mouvement	23	4,03	enfants	106	3,63	enfants	105	2,73	parité	19	3,25
problèmes	12	3,86	hommes	201	3,26	alphabétisation	9	2,41	postes	22	2,69
main-d'oeuvre	9	2,54	aujourd'hui	16	2,45	terre	10	2,28	société	47	2,64
travailleurs	15	2,34	noirs	15	2,38	rues	9	2,11	sexes	20	2,62
ouvriers	12	2,32	jour	15	2,29	ans	41	2,10	accès	22	2,53
longtemps	9	2,25	afrique	12	2,28	travaillent	14	2,02	droits	67	2,49
aujourd'hui	15	2,16	enceintes	12	2,17	familles	11	1,95	études	17	2,22
paysans	9	2,11	âge	14	1,94	développement	14	1,92	question	26	2,16
lieu	11	2,03	mères	10	1,90	taux	10	1,89	selon	24	2,06
population	17	1,99	majorité	17	1,79	main-d'oeuvre	8	1,78	france	33	2,01

Figure 2. Graphe les cooccurrents forts de femmes *sur 4 périodes de 5 ans du* Monde Diplomatique

on pourra notamment favoriser le retour vers l'ensemble des phrases caractéristiques de ces périodes, celles où ces cooccurrents forts sont concentrés.

Ainsi pour la période 1980-84 : (Patrick Tissier, « LA RELANCE DE L'ÉCONOMIE », Octobre 1983)

> Les primes liées au travail aux pièces renforcent les disparités entre travailleurs à la production et travailleurs auxiliaires : par exemple, dans l' industrie lourde, la majorité de ces derniers sont des femmes moins payées que les hommes, surtout engagés dans la production.

Ou pour 1996-2000 : (Olfa Lamloun, « DISCOURS MODERNISATEUR POUR RÉGIME RÉPRESSIF », Juin 1998)

> Pour autant, l'hymne au Sâna'al Tahauil occulte le rôle des tunisiennes et la mobilisation des féministes, qu'il s'agisse du club Tahar Haddad, de la commission d'études de la condition des femmes travailleuses au sein de l'Union Générale des Travailleurs Tunisiens, de la commission de défense des droits des femmes de la Ligue Tunisienne des Droits de l'Homme, ou encore de l'Association des Femmes Démocrates.

On comprendra qu'il s'agit ici d'une pratique essentiellement pré-exploratoire, qui n'a d'autre objet que de soulever des questions en faisant, par principe, monter les faits saillants de la structure elle-même. C'est ainsi que le choix des pivots peut être lui-même plus ou moins automatique. Rappelons qu'en l'occurrence nous avons porté notre attention sur *femmes* en raison de sa position deux fois polaire (en micro- et en macro-distributions).

On peut bien sûr objecter qu'il est dans une certaine mesure normal que le vocabulaire cooccurrent de *femmes*, qui représente en masse, selon les paramètres choisis, 110'000 mots environ soit 0,65 % de l'ensemble du corpus, suive les tendances diachroniques d'ensemble. Cela n'enlève rien à l'utilité de ce graphe, qui nous informe sur la contribution particulière de *femmes* à la macro-distribution diachronique d'ensemble (l'un ne pouvant en aucun cas être indépendant de l'autre).

Mais cela exige de compléter l'information par un autre aperçu d'ensemble, où l'évolution des grandes masses aura été compensée de manière à voir comment, compte tenu même des emplois croissants et décroissants, parmi ce vocabulaire en mouvement, *femmes* modifie la « sélection » de ses cooccurrents. Pour cela, il nous faut considérer, pour chaque cooccurrent dans chaque partie du corpus (année), la proportion de l'ensemble de ses emplois qui appartiennent au contexte de *femmes*. Cette proportion peut être alors appliquée à une matrice « idéale » de 21 parties égales en volume, où les cooccurrents se répartiraient équitablement.

Ainsi, pour *travail* qui a 8806 occurrences, la base de calcul pour une partie sera de 8806/21, soit 419,3. Dans l'année 1987, on constate que 7 des 273 occurrences effectives de *travail* sont en contexte avec *femmes*, soit une proportion de 0,0256. En appliquant 0,0256 à 419,3, on établit la cooccurrence compensée de *femmes* et de *travail* dans l'année 1987, qui est donc de 10,75 (que l'on arrondit à 11).

L'AFC de la matrice des données ainsi compensées offre ce graphe :

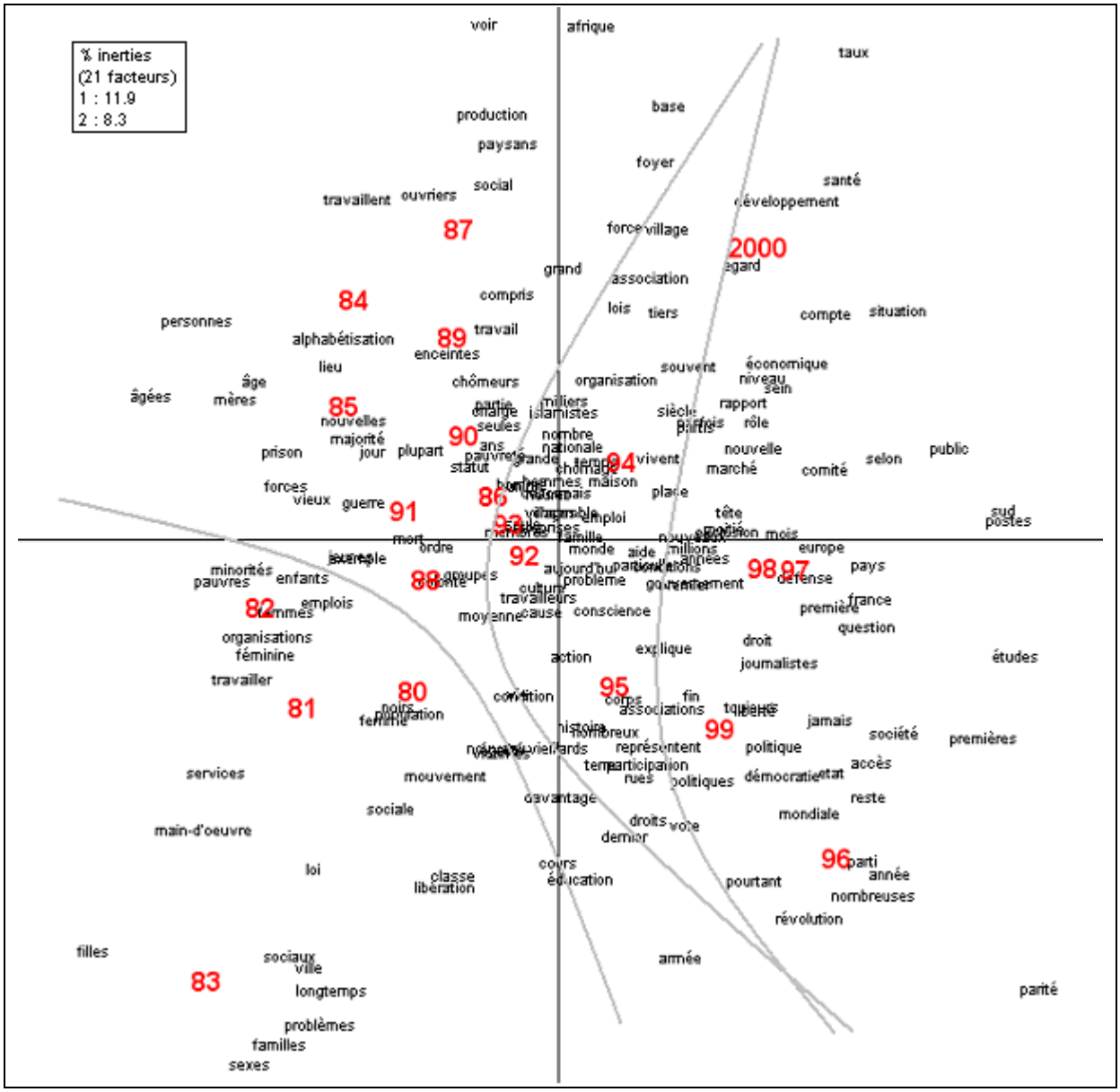


Figure 3.
Graphe macro-distributionnel des cooccurrents [matrice compensée] de femmes. *186 formes lexicales.*

La différence est assez minime. L'orientation diachronique est confirmée, voire renforcée par la restriction du cadre d'étude. L'évolution de ses cooccurrents est relativement autonome par rapport à celle des grandes masses du vocabulaire du *Monde Diplomatique*. On peut dès lors parler de diachronie propre de l'item lexical dans la temporalité du corpus. Il y a bien, dans ce secteur du discours politico-institutionnel, une modification des valeurs de l'item, telles que l'attestent les environnements différents.

Savoir quelle est cette modification relève d'une étude minutieuse des contextes dont il n'est pas question de rendre compte dans les limites ici imparties. Cette étude est facilitée par la collecte ergonomique qui en sera faite par simple click sur un cooccurrent indiqué sur la carte.

Par contraste avec *femmes*, nous présenterons plus sommairement encore le cas de *réalité(s)*, qui a retenu nos premières attentions du fait qu'il joue (avec d'autres), dans le discours politico-institutionnel, un rôle certes obscur, mais néanmoins irremplaçable, de vecteur de l'implicite socio-économique (plus encore que *loi(s)* ou *contrainte(s)*, plus patents).

Comment le discours d'un journal marqué à gauche (et cela de façon croissante entre 1980 et 2000, au point de le faire classer aujourd'hui dans la gauche « radicale »), se tire-t-il de l'emploi délicat (et *a priori* peu auto-surveillé) de ces formes ? On compte 3670 occurrences du pluriel et 751 du singulier.

On note tout d'abord que cet emploi tend à s'élever, comme le suggère l'histogramme des écarts-réduits à l'équi-distribution par année :

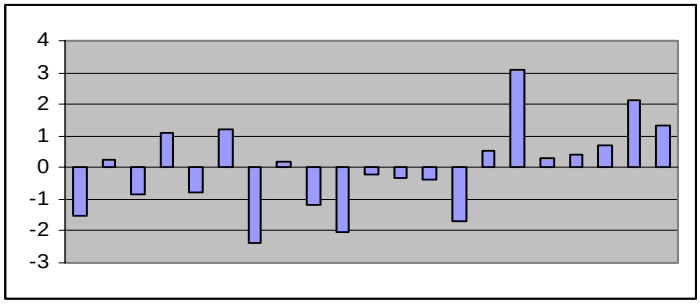


Figure 4. Histogramme des écarts-réduits à l'équi-distribution de réalité(s) dans le corpus LMD

Pour les 7 dernières années (94 à 2000), l'occurrence constatée est 1635, contre 1511 dans l'hypothèse d'équi-répartition : l'écart-réduit est de 3,2. Par ailleurs, la proportion des emplois au singulier s'élève assez sensiblement à partir de 1990 (de 15 à 18 %).

Surtout, si l'on compare sur des cartes « muettes » (où seules les années sont indiquées, et non les formes recensées), les distributions diachroniques de ses cooccurrents, « brute » à gauche,

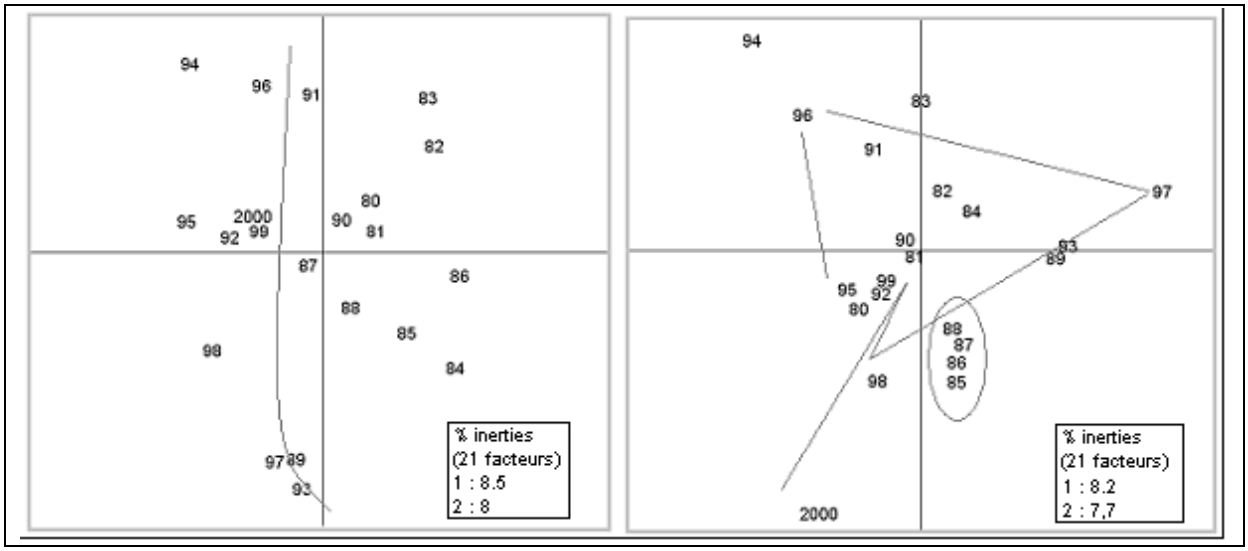


Figure 5. Graphes macro-distributionnels des cooccurrents de réalités(s).
À gauche, sans compensation ; à droite, avec compensatio.

« compensée » à droite, on relève que l'orientation encore sensible en « brut » (c'est-à-dire lorsqu'elle récupère une part notable de l'inertie de la distribution totale), se dissipe presque entièrement, ou du moins se fragmente en « compensé » (on repère encore quelques séquences d'années, mais on voit surtout de grands et brusques écarts). La participation de *réalité(s)* à l'évolution diachronique des masses du vocabulaire n'est donc pas clairement distincte de sa propre évolution distributionnelle. En d'autres termes, ils ne jouent pas un rôle cohésif dans le discours du *Monde Diplomatique* comme diachronie signifiante globale.
Cela n'implique pas que l'étude diachronique de leurs collocations soit sans intérêt, au contraire, puisqu'il y a sans doute moins de maîtrise idéologique sur l'emploi de ces formes : donc, à tout le moins, une attention spécifique à leur porter (surtout, compte tenu de l'augmentation significative de leur occurrence).

3. Application à une controverse d'attribution : Corneille/Molière.

Nous avons montré *supra* un exemple d'application purement pré-exploratoire.

Cette méthode d'analyse séquencée de la micro-distribution peut aussi être convoquée pour dialoguer avec une hypothèse de travail constituée. Dans son argumentation destinée à soutenir que Corneille a réellement écrit 16 des pièces attribuées à Molière, D. Labbé (2003 : 111-116) développe notamment (en substance) que les collocations d'items essentiels (comme *amour* par exemple) sont proches chez Corneille et dans ces 16 pièces.

Il nous semble possible de vérifier cette assertion de manière assez objective et précise, en employant l'analyse séquencée de la micro-distribution, dans une perspective non plus diachronique, mais simplement contrastive. Relevons, au fil des 67 pièces du corpus, les contextes de pivots lexicaux mentionnés par D. Labbé, et distribuons-les dans une matrice de 67 colonnes. Prenons l'exemple des vocables *amour* et *aimer* (cumulés), dans les limites de la phrase (25 mots à droite et à gauche)[2], pour lesquels nous avons, comme exposé *supra*, sélectionné les 201 premiers cooccurrents (en cooccurrence brute).

Avec 67 sections de corpus, il ne peut être question de visualiser en même temps la distribution des items et la distribution dans les pièces. Nous montrerons donc seulement la carte « muette », avec les seules positions des pièces. Dans l'environnement ®Astartex, ce sont des codes couleurs qui permettent de visualiser les différentes classes qui nous intéressent. Sur ce papier, ce sont des conventions graphiques. Les titres (abrégés) des pièces de Molière sont soulignés. Ceux des pièces non attribuées à Corneille sont estompés en gris. Ceux des comédies de Corneille sont en italiques. Nous avons en outre marqué de l'astérisque les pièces de Molière antérieures à 1662.

Cette distribution (fig. 6, p. suiv.) ne prouve certes rien, mais elle apporte des éléments considérables pour indiquer au contraire que les vocables *amour* et *aimer*, comme la plupart d'ailleurs des vocables significatifs du corpus, sont traités de façon fort différente par les deux auteurs, à l'exception des cas de *Dom Garcie de Navarre*, proche du genre de la comédie héroïque et depuis longtemps connue par les spécialistes pour avoir été très fortement influencé par Corneille, et des premières pièces de Molière, qu'elles soient ou non réattribuées à Corneille, et qui se rapprochent en effet dans cette distribution des comédies de ce dernier.

[2] Le corpus est celui de D.Labbé, lemmatisé par ses soins.

Ce n'est pas le critère de la réattribution supposée à Corneille (titres en noir *vs* titres en gris) qui justifie l'éloignement des comédies de Molière par rapport au noyau de celles de Corneille (en haut du graphe), mais la diachronie.

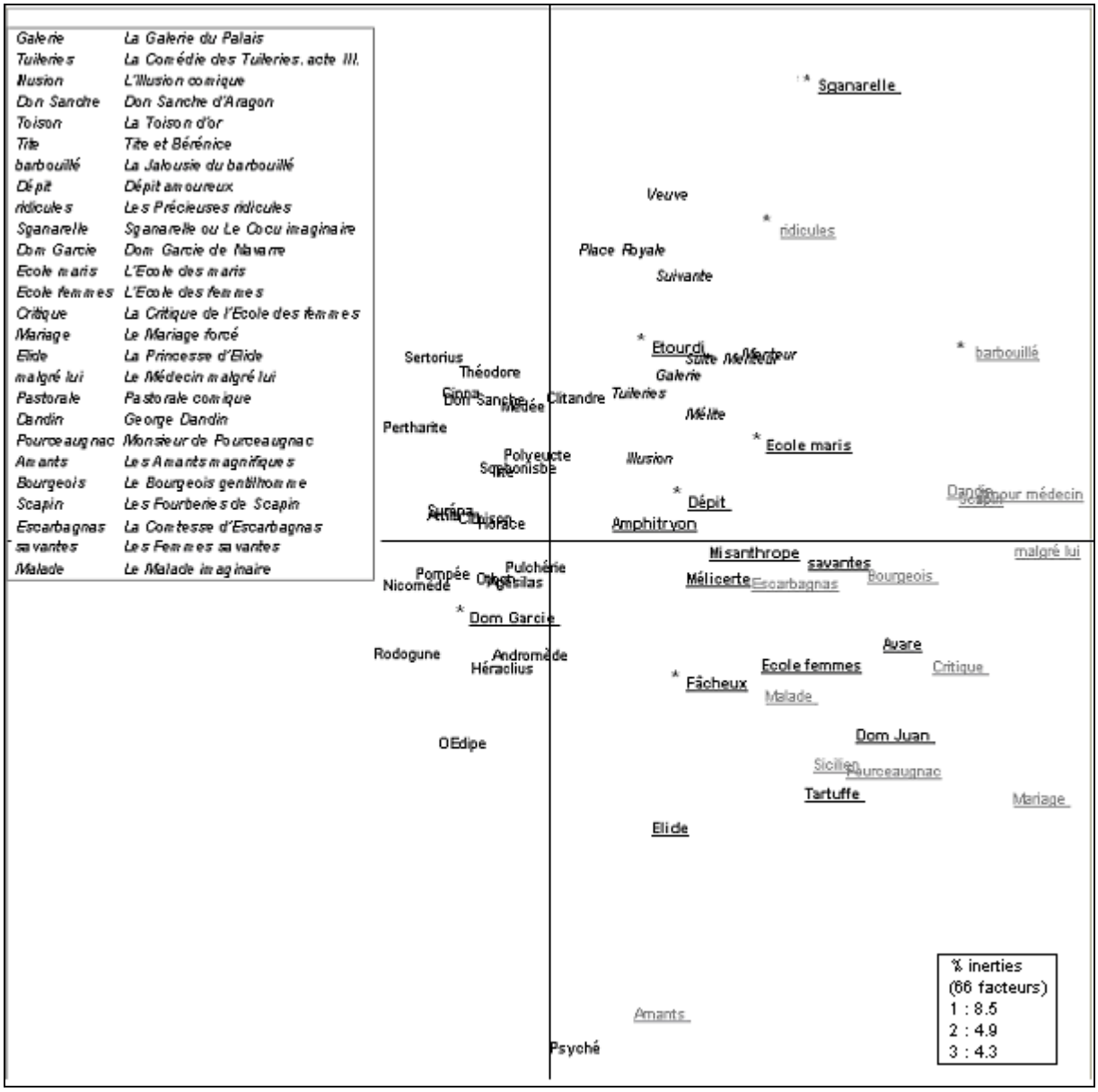


Figure 6. Distribution des co-occurrents de amour *et* aimer *dans le corpus Corneille/Molière*

En revanche, si l'on se libère de l'optique inférentielle pour se replacer dans une perspective exploratoire, on peut maintenant observer, à partir de la cooccurrence séquencée, d'authentiques variations quant au comportement de certains vocables dans cette comparaison. C'est ainsi que, pour le couple *père-fils*, les proximités repérées par D. Labbé semblent plus actives, ainsi que le montrent les positions respectives des pièces de Molière « réattribuées » (+) ou non (-) par rapport au sous-nuage Corneille, ainsi que la position des 2 *Menteurs*. On peut en déduire, avec toute la prudence de mise, que certains vocables sont bel et bien responsables, plus que d'autres, des proximités lexicales fines, et non massives, entre nos deux auteurs (fig. 7, p. suiv.).

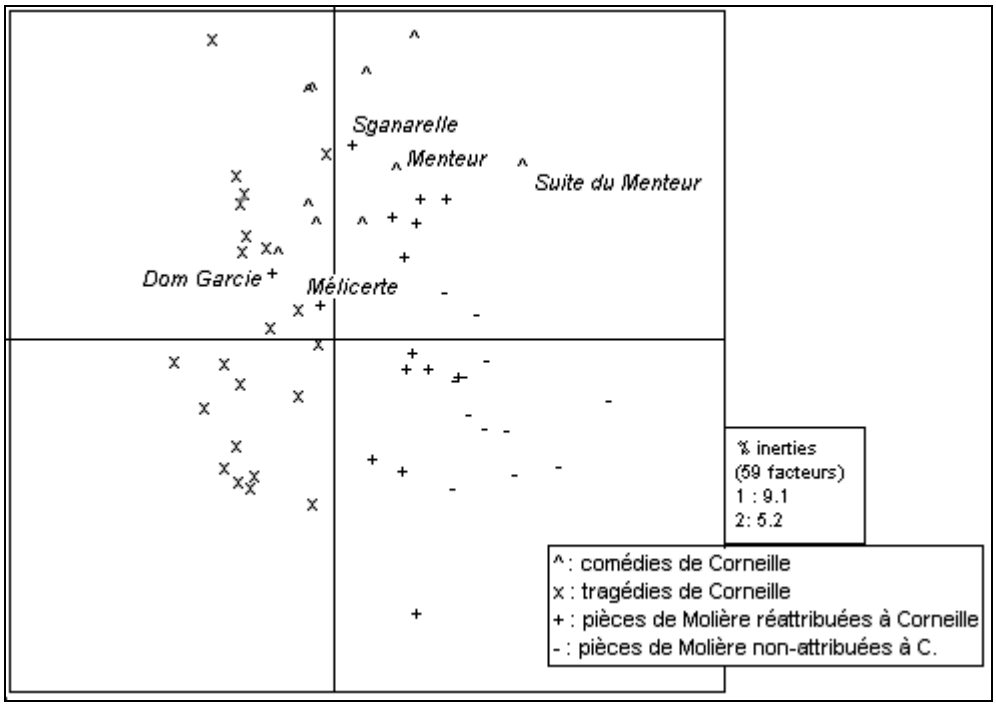


Figure 7. Distribution des co-occurrents de père *et de* fils *dans le corpus Corneille/Molière*

Il reste un test à produire à partir de ces collectes de données. Pour *amour/*aimer, comme pour *père/*fils, nous disposons de matrices de distribution des cooccurrences, dont nous pouvons regrouper les données selon une classification pertinente des pièces du corpus. Nous créerons 4 classes : les comédies de Corneille (CC), les autres pièces de Corneille (CT), les pièces de Molière réattribuées à Corneille par D. Labbé (MA), celles qu'il juge douteuses ou n'attribue pas du tout (MN). Nous créerons également 3 classes mixtes, où nous regrouperons CC et CT en C, MA et MN en M, C et MA en CMA.

Pour chacune des 7 classes, nous attribuerons un rang à chacun des cooccurrents de *amour/aimer*, selon sa cooccurrence brute. Puis, nous calculerons la corrélation de ces rangements (indice de Spearman[3]). Bien sûr, toutes les corrélations seront positives puisque l'ordre décroissant de cooccurrence dépend très largement de l'occurrence totale. Ainsi, *faire* occupe-t-il toujours et partout le 1er rang et *voir* ne descend-il jamais au-delà du 4e. Cependant, les variations de l'indice de Spearman sont tout à fait éclairantes. Une seule corrélation s'élève significativement au-dessus des autres : celle qui résulte de la comparaison entre les deux sous-corpus de Molière, qui montrent ainsi leur très grande homogénéité de ce point de vue. Les autres corrélations sont très voisines les unes des autres, et notamment deux qui sont décisives du point de vue de l'hypothèse explorée : entre les comédies de Molière réattribuées à Corneille et (1) les comédies de Corneille (2) l'ensemble de Corneille.

	MA/C	M/C	MA/MN	M/CC	CC/CT	MA/CC
amour+aimer	0,5037	0,4880	0,7519	0,5199	0,4915	0,5056
père+fils	0,4559	0,4132	0,7631	0,5652	0,4175	0,5766

Figure 8. Corrélation (Spearman) des rangs des plus forts occurrents dans les sous-corpus indiqués

[3] L'indice de Spearman oscille entre +1 – identité de rangement -- et +1 – rangement inverse. 0 indique une corrélation nulle. L'interprétation de l'indice est développée par Muller (1992-A : 153 ssqq). Nous nous contentons ici de comparer les indices pour des couplages donnés de sous-corpus.

Plus précis encore : ce tableau nous montre bien que, contrairement à *amour+aimer*, *père+fils* présentent une meilleure corrélation entre MA et CC, qu'entre MA et C, ce qui en même temps confirme le graphe n°6 (parenté relative de MA et des comédies de Corneille du point de vue de l'emploi de *père+fils*), mais aussi *relativise* cette parenté : elle reste très en-deça de celle constatée, pour les mêmes items, entre les deux sous-corpus de Molière. Avec la corrélation MA/CC, nous excluons le facteur *genre* : il n'y a plus, si nous acceptons cette simplification, que le facteur *auteur*…

4. Conclusions

L'analyse séquencée de la cooccurrence lexicale permet de combiner les analyses micro-distributionnelle et macro-distributionnelle, au profit d'une exploration plus poussée des corpus textuels.

Elle permet d'affiner notamment l'étude diachronique et/ou contrastive de la signification d'un item, ou d'un groupe d'items, telle qu'elle se développe dans les limites d'un corpus constitué, à partir de la norme endogène de celui-ci. Les items étudiés étant eux-mêmes extraits du vocabulaire grâce à des processus endogènes (micro- et macro-distributions globales dans le corpus), on a toutes les chances, tout en conservant un niveau élevé d'objectivation, d'être orienté rapidement et ergonomiquement, dans le cadre d'un hypertexte expert, vers des phénomènes hautement signifiants.

Cette démarche permet aussi de dialoguer utilement avec des hypothèses fortes, dans l'objectif de les vérifier certes, mais surtout de les raffiner pour le meilleur profit de l'analyse de textes.

5. Références bibliographiques

Coll. (2001). *Le Monde diplomatique 1980-2000 : 21 ans d'archives*. CEDROM-SNI, Québec.

Labbé D. (2003). *Corneille dans l'ombre de Molière*. Les Impressions nouvelles.

Muller Ch. (1992a). *Initiation aux méthodes de la statistique linguistique*. Champion.

Muller Ch. (1992b). *Principes et méthodes de statistique lexicale*. Champion.

Schepens P., Lelu A. et Viprey J.-M. (2004, à paraître). *Essais d'analyse de discours sur des corpus d'entretiens cliniques*. Les Belles-Lettres.

Viprey J.-M. (1997). *Dynamique du vocabulaire des Fleurs du mal*. Champion.

Viprey J.-M. (2000). « Hypertexte de corpus littéraire : cartographie et statistique multidimensionnelle ». In *Actes des JADT 2000* : 535-538.

Viprey J.-M. (2002a). « Dynamisation de l'analyse micro-distributionnelle des corpus textuels ». In *Actes des JADT 2002* : 779-790.

Viprey J.-M. (2002b). *Analyses textuelles et hypertextuelles des* Fleurs du mal. Champion.

Annexes :

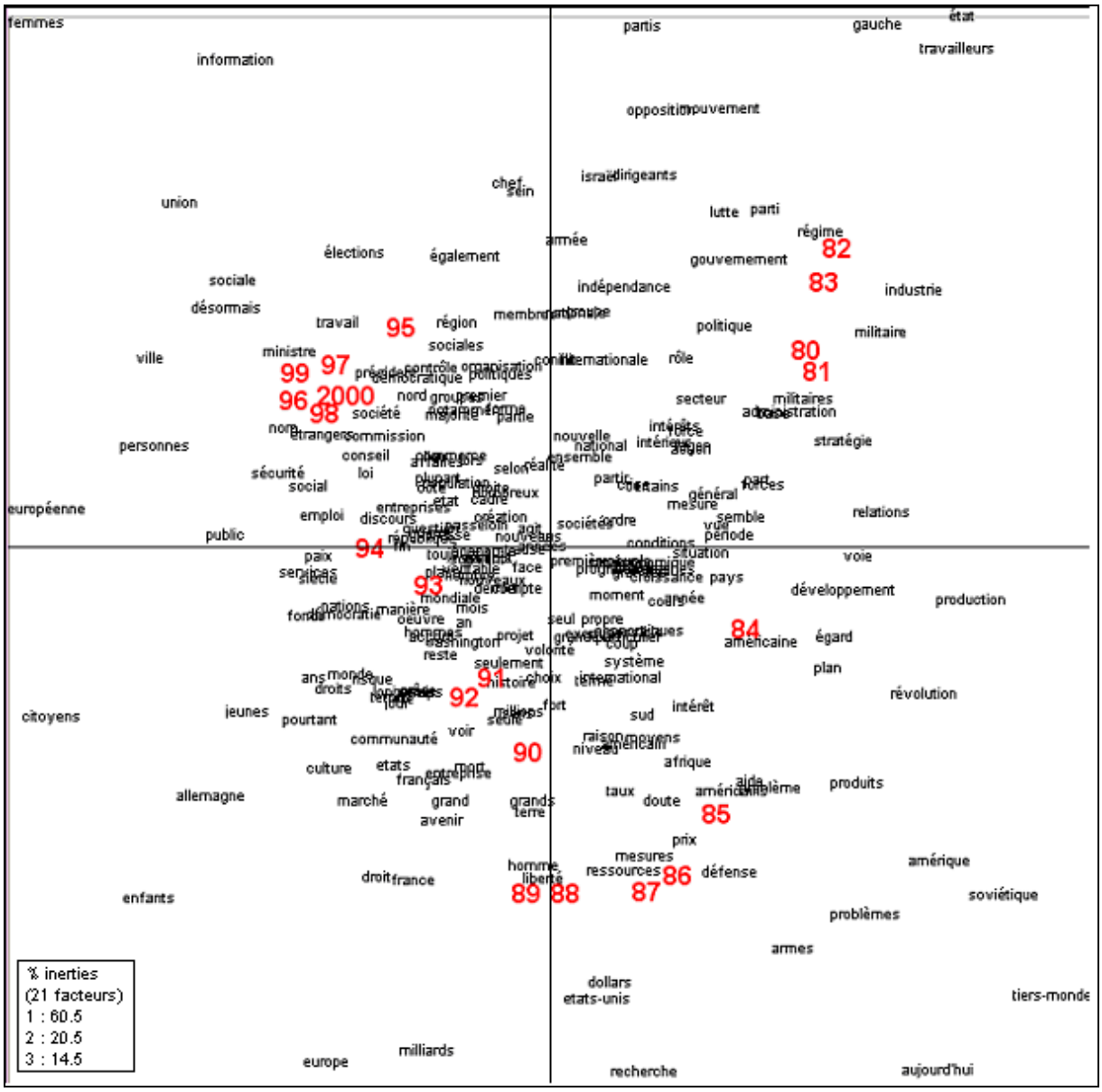


Figure A1. Graphe macro-distributionnel. 240 formes lexicales du corpus Monde Diplomatique

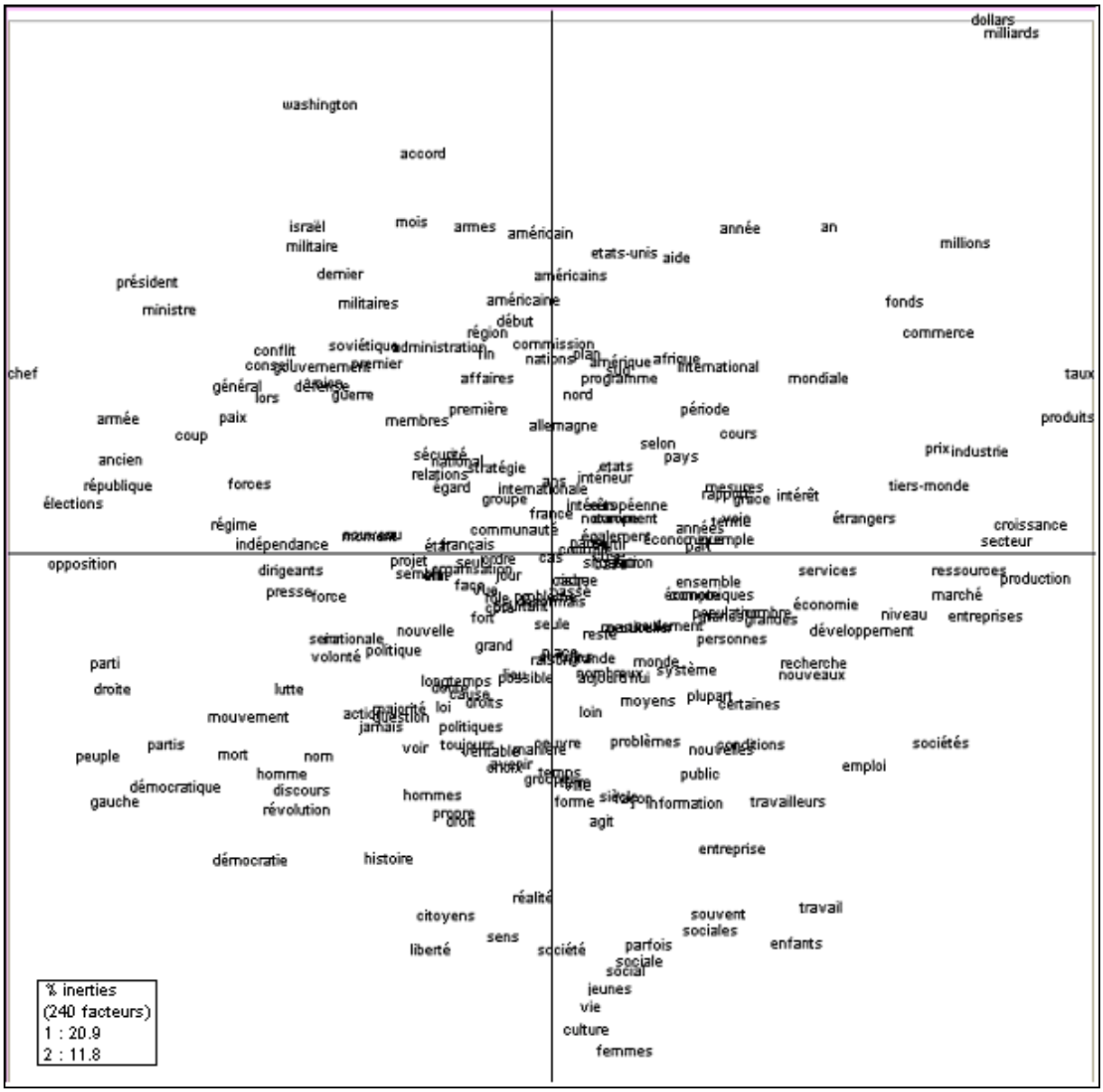


Figure A2. Graphe micro-distributionnel. 240 formes lexicales du corpus Monde Diplomatique

SANDOH, un outil pour analyser des textes hétérogènes

Vo Trung Hung

GETA-CLIPS-IMAG, Institut National Polytechnique de Grenoble
385, rue de la Bibliothèque – BP 53 – 38041 Grenoble Cedex 9 – France
Tél : +33 4 76 51 48 17 – Fax : +33 4 76 51 44 05
Hung.Vo-Trung@imag.fr

Abstract

The knowledge of the language and the coding used for the texts helps us to correctly carry out treatments on these texts as the exact search for information, the checking of orthography, the visualizing of information, the exchange of information between different systems, etc. In this article, we present the construction of a tool to analyse a heterogeneous text by method diagnoses N-grams and method of progressive segmentation to segment a heterogeneous text in homogeneous zones. This tool makes it possible to analyse a text to identify the language and coding used in this text. For the homogeneous text, the result is a couple <language, coding>, for the heterogeneous text, the result is a table of the zones with the language and corresponding coding {zone-1, language-1, coding-1}, {zone-2, language-2, coding-2}, …, {zone-n, language-n, coding-n}.

Résumé

La connaissance de la langue et du codage utilisé dans les textes aide à réaliser correctement des traitements sur ces textes comme la recherche exacte d'informations, la vérification d'orthographe dans la langue utilisée, l'affichage d'informations avec le bon jeu de caractères correspondant au codage utilisé, l'échange d'informations entre systèmes différents, etc. Dans cet article, nous présentons la construction d'un outil pour analyser un texte hétérogène selon méthode du diagnostic n-grammes et méthode de la segmentation progressive pour segmenter un texte hétérogène en zones homogènes. Le résultat de l'analyse est le nom d'un couple <langue, codage> si ce document est homogène, sinon, les zones et le couple <langue, codage> utilisé dans chaque zone {zone-1, langue-1, codage-1}, {zone-2, langue-2, codage-2}, …, {zone-n, langue-n, codage-n}.

Mots-clés : multilinguisme, texte multilingue, texte homogène, texte hétérogène, traitement de la langue naturelle, segmentation, classification, multilinguïsation.

1. Introduction

Actuellement, on peut traiter tous les systèmes d'écriture par l'ordinateur, mais on doit utiliser souvent plusieurs systèmes de codage pour présenter un système d'écriture (par exemple : ISO 8859-1, CP1252, CP 850, Macintosh, … pour le français ; ISO 8859-1, CP1252, TCVN, VNI, VPS, … pour le vietnamien ; ISO 8859-5, CP 1251, CP 886, KOI 8, Mac Cyrillic, … pour le russe) et c'est difficile de traiter ces données. Malgré Unicode, les utilisateurs rencontrent beaucoup de problèmes avec des fichiers contenant des couples de langues et de codages très différents. En pratique, on a souvent l'expérience frustrante de vouloir ouvrir une page Web ou un fichier, d'essayer tous les codages proposés, et de ne pas arriver à la lire, sans parler de l'éditer.

La connaissance de la langue et du codage utilisé dans les textes aide à réaliser correctement des traitements sur ces textes comme la recherche exacte d'informations, la vérification d'orthographe dans la langue utilisée, l'affichage d'informations avec le bon jeu de caractères correspondant au codage utilisé, l'échange d'informations entre systèmes différents. En parti-

culier, avec les documents hétérogènes (écrits en plusieurs langues et plusieurs codages), l'identification et la segmentation nous permettent traiter correctement pour chaque zone de langue et de codage différent. Par exemple, à chaque zone nous appliquons un dictionnaire propre pour vérifier les fautes d'orthographe.

Nous présentons ici l'outil SANDOH (Système d'ANalyse des DOcuments Hétérogènes) que nous avons construit pour analyser un texte hétérogène en zones homogènes. Le résultat du système est le nom d'un couple <langue, codage> si le document est homogène, sinon, les zones et le couple <langue, codage> utilisé dans chaque zone.

2. État de l'art

Cette section aborde quelques systèmes récents d'identification automatique la langue et le codage du texte homogène (écrit en seule langue et seul codage). La liste n'est pas exhaustive.

2.1. *SILC (Système d'Identification de la Langue et du Codage)*

Le logiciel SILC est développé au laboratoire RALI de l'Université de Montréal principalement par Pierre Plamodon. SILC reconnaît 28 langues encodées avec en moyenne trois encodages par langue. En totalité, il peut identifier 73 couples <L, C> différents. SILC utilise associer les unigrammes et trigrammes pour évaluer des textes. C'est un bon outil diagnostique des textes homogènes (Bouffard, 2002 ; Vo-Trung, 2003).

2.2. *TextCat*

TextCat est un système développé par Gertjan Van Noord de l'Université de Groningen (http ://odur.let.rug.nl/~vannoord/TextCat/). Ce système implante une méthode basée sur les ngrammes telle que décrite par Wiliam Cavnar et John Trenke. Il reconnaît en tous 70 langues et un total de 77 paires <L, C>.

2.3. *Xerox Mltt Language Guesser*

Ce système a été développé par Xerox et est commercialisé par Inxight Software. Il reconnaît 47 langues chacune encodée avec en moyenne 2 jeux de caractères pour un total d'exactement 93 couples <L, C> (www.xrce.xerox.com/research/mltt/tools/guesser.html).

2.4. *Lextek Language Identifier*

Lextek est un système de l'identification automatique de la langue et du codage de Lextek International. Ce système peut diagnostiquer pour 260 couples <L, C> différents (www.languageidentifie.com).

3. Objectif

Ces systèmes s'appliquent principalement à l'identification des fichiers homogènes en format texte ou HTML. L'exactitude du diagnostic varie beaucoup suivant la longueur du texte analysé. Avec les textes de plus de 50 lettres, le diagnostic est assez bon mais la qualité de diagnostic est trop mal avec les textes inférieurs à 30 lettres (Vo-Trung, 2003). Un problème posé qu'il faut chercher une nouvelle solution pour analyser des textes très courts.

Notre but est de proposer un système d'analyse des textes hétérogènes en zones homogènes sur la base du système d'identification d'un texte homogène. Nous proposons à utiliser des dictionnaires monolingues pour diagnostiquer des textes très courts.

Dans la partie suivante, nous présentons l'architecture du système et les méthodes utilisées pour ce système.

4. Système d'analyse d'un document hétérogène

Dans les systèmes multilingues, on doit souvent travailler avec des données hétérogènes qui sont écrites en plusieurs langues et peut-être sous des codages différents. Par exemple, les documents français contiennent des citations anglaises ou bien les dictionnaires contiennent plusieurs articles écrits dans des langues différentes. La connaissance de la langue et du codage utilisé dans chaque zone de ces textes aide à réaliser correctement des traitements à la suite.

4.1. *Architecture du système*

Le système d'analyse d'un texte hétérogène est organisé :

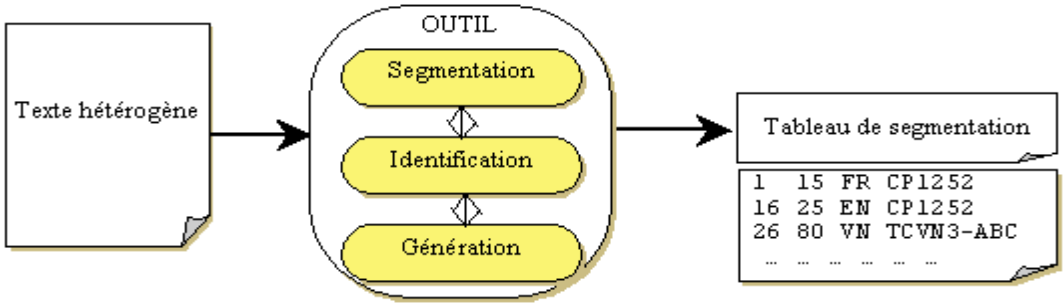


Figure 1. Outil d'analyse d'un texte hétérogène en zones homogènes

Cet outil a trois fonctions principales : la segmentation, l'identification et la génération.

- Le module de segmentation permet de segmenter un texte en zones à évaluer. Chaque zone est définie par la première et la dernière position de lettres de la zone. Par exemple, la zone 1 est définie par (p_{1z1}, p_{2z1}), la zone 2 par (p_{1z2}, p_{2z2}),…, la zone n par (p_{1zn}, p_{2zn}).
- Le module d'identification permet de diagnostiquer la zone actuelle pour évaluer si elle est hétérogène ou homogène. Si elle est homogène, il renvoie le résultat (zone, langue, codage) au module de génération et passe à une autre zone. Sinon, il renvoie cette zone au module de segmentation qui la segmentera en zones plus petites.
- Le module de génération permet d'écrire le résultat obtenu dans un fichier. Ce résultat est une table, chaque ligne de cette table présente une zone (la position de la première lettre et de la dernière lettre) avec la langue et le codage utilisé sur cette zone.

4.2. *Segmentation*

Pour segmenter un texte, nous utilisons une technique de segmentation progressive basée sur les signes de ponctuation. Normalement, les parties écrites dans des langues différentes sont séparées par des signes de ponctuation comme le point, le tiret, les deux-points, la parenthèse, le guillemet, le point d'exclamation, le point-virgule, etc. L'idée ici est qu'après avoir évalué une zone, si elle est hétérogène, on continue de séparer cette zone en zones plus petites pour l'évaluer. Mais il faut s'assurer que ces zones ne seront pas trop courtes pour l'identification.

D'abord, nous évaluons chaque paragraphe (la terminaison de paragraphe est un signe de retour à la ligne) pour diagnostiquer s'il est homogène ou hétérogène. S'il est homogène, on

passe à un autre paragraphe, sinon, il faut découper ce paragraphe en deux zones. Il suffit de séparer le texte au milieu et ensuite de tester ce qui arrive lorsque l'on se déplace d'un mot vers la gauche ou vers la droite jusqu'à obtenir que chaque zone contienne un ensemble de phrases. Nous continuons d'évaluer et de séparer cette zone en zones plus petites et à évaluer jusqu'à obtenir une zone homogène.

4.3. *Identification de la langue et du codage d'une zone homogène*

4.3.1. *Architecture générale du système d'identification de la langue et du codage*

Il y a en général deux phases pour identifier la langue et le codage : la première phase crée un modèle du couple <langue, codage> (seulement une fois pour chaque couple) et la deuxième phase identifie la langue et le codage du texte à analyser.

L'identification de la langue et du codage est réalisée comme le montre le schéma ci-dessous (Russell, 2002) :

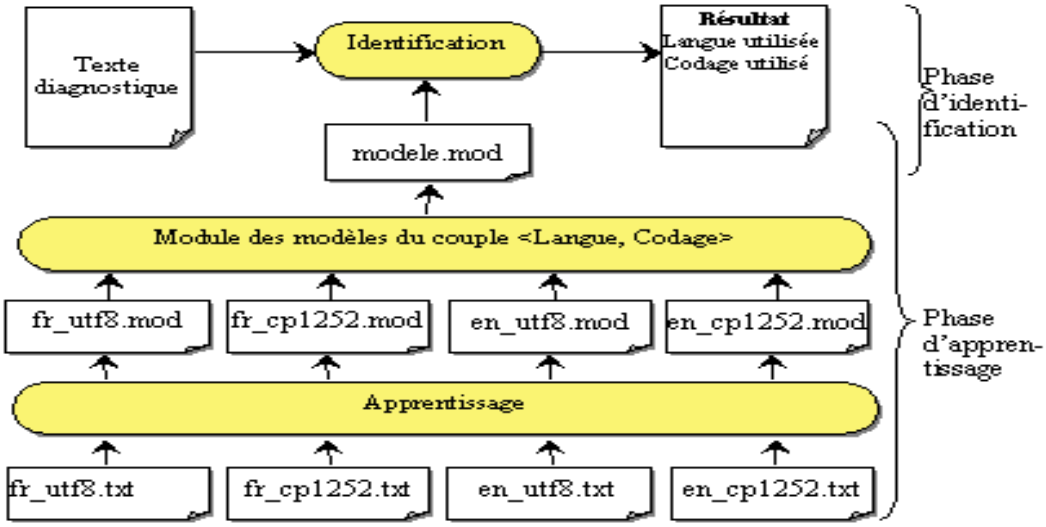


Figure 2. Architecture générale d'un identifieur de langue et de codage

La phase d'apprentissage est construite sur la base d'un modèle statistique. Au début, on doit avoir des fichiers annotés par des balises <langue, codage>. Le module d'apprentissage crée ensuite les « modèles » correspondant à chaque couple <langue, codage>. Ces modèles sont construits à partir de la fréquence des séquences qu'on a comptées dans le fichier d'apprentissage. Puis on réalise la synthèse de ces modèles pour réunir les fichiers obtenus en un fichier unique qui constitue le modèle du couple <langue, codage>. Ce fichier est utilisé pour identifier les autres textes dans le futur.

La phase d'identification permet de déterminer en quelle langue est écrit un texte et avec quel codage. Elle utilise la même méthode en comparant des segments dans le texte à analyser avec des séquences du modèle de langues pour évaluer le texte à analyser.

4.3.2. *Méthode n-grammes*

On calcule d'abord la fréquence des séquences de ***n caractères*** dans un grand fichier d'apprentissage. L'idée ici est de détecter des motifs récurrents dans des suites d'observations (Cavnar et Trenkle, 1994). Par exemple, les mots anglais se terminant par ***-ck*** sont plus nombreux que les mots français avec la même terminaison. Par contre, les mots français se termi-

nant par ***-ez*** sont plus nombreux que les mots anglais présentant la même fin. Ces motifs récurrents peuvent avoir une longueur variable.

L'algorithme d'apprentissage est un algorithme itératif sur la base d'apparition des éléments des séquences n lettres dans le fichier d'apprentissage (Beesley, 1998). Au cours de ces itérations, les segmentations du corpus évoluent, faisant émerger les séquences d'observation les plus typiques.

Après apprentissage, un modèle du couple <langue, codage> est créé, contenant les séquences sélectionnées les plus probables et leur vraisemblance.

Par exemple, avec le texte d'apprentissage « *les chiens et les chats sont des animaux* », on va obtenir les modèles de langage n-grammes suivants :

Tableau 1.
Modèles de langage avec leurs séquences et leurs fréquences

Unigrammes		**Bigrammes**		**Trigrammes**	
Séq	**Fréq**	**Séq**	**Fréq**	**Séq**	**Fréq**
_	7	s_	5	es_	3
s	6	es	3	les	2
e	5	le	2	s_c	2
a	3	_c	2		
n	3	ch	2		
t	3				

La phase d'identification consiste à calculer la probabilité sur le texte à analyser et sur les modèles du couple <langue, codage>. On choisit alors la langue et le codage correspondant au modèle du couple <langue, codage> qui est le plus probable.

On peut calculer le taux de certitude qui caractérise une langue et un codage correspondant à chaque modèle du couple <langue, codage> par la formule (Manning et Schutze, 1999) :

$$l = \sum_{i=1}^{N} P_i \qquad (1)$$

Avec :

$$P_i = \begin{cases} \log\left(\frac{f(w_i)}{N} x \frac{F(w_i)}{M}\right) & \text{, si } w_i \in \text{ texte} \\ 0 & \text{, si } w_i \notin \text{ texte} \end{cases}$$

N = nombre de séquences dans le modèle du couple <langue, codage> actuel

$f(w_i)$ = fréquence de la séquence w_i dans le modèle du couple <langue, codage> actuel

$F(w_i)$ = fréquence de la séquence w_i dans le texte à analyser (si $w_i \in$ texte)

M = nombre de séquences dans le texte à analyser

On conclut sur la langue et le codage appliqué sur le texte en choisissant le modèle du couple <langue, codage> qui a le taux de certitude maximum.

C'est une méthode utilisable pour tout couple <langue, codage>. La construction du modèle et l'évaluation du texte ne dépendent pas de la langue et du codage, elles consistent à analyser les segmentations des textes.

On peut changer la variable n pour obtenir des modèles différents tels qu'unigramme (n=1), bigrammes (n=2), trigrammes (n=3), tétragrammes (n=4),… et on peut aussi associer des modèles pour augmenter l'exactitude du système.

Dans notre système, nous avons utilisé un modèle du couple <langue, codage> sur des n-grammes avec n variant de 1 à 4, et chaque modèle contenant les 400 séquences de plus haute fréquence.

4.3.3. *Résultat obtenu*

Nous avons développé un outil d'identification d'un texte homogène selon la méthode ci-dessus pour plusieurs langues et codages différents. Le système actuel permet de diagnostiquer 53 langues comme l'anglais, le français, le vietnamien, le chinois, le russe, l'arabe, etc. Nous pouvons facilement ajouter d'autres langues par fourniture des fichiers d'apprentissage.

Le résultat obtenu est un tableau de scores correspondant aux modèles du couple <langue, codage> analysés et nous choisissons la langue et le codage selon le modèle du couple <langue, codage> de plus grand score.

Voici un exemple de résultat sur un texte français « *les chiens et les chats sont des animaux* » et un texte chinois 地定空虚混沌。渊面黑暗。神的灵运行在水面上 avec les 10 scores les plus hauts :

Texte français			**Texte chinois**	
Langue-codage	Score		Langue-codage	Score
Français	255.8934		Chinois-gb2312	344.3639
Roman	236.9501		Chinois-big5	234.7359
Écossais	231.5291		Coréen	222.3591
Espagnol	211.8286		Arabe-windows1256	170.5808
Latin	195.9060		Tamoul	169.7671
Slovene-iso8859_2	195.1329		Arabe-iso8859_6	140.7884
Irlandais	183.8261		Ukrainien-koi8_r	135.7941
Quechua	167.9354		Japonais-shift_jis	128.0966
Slovaque-windows1250	167.8815		Thaï	119.0824
…	…		…	…

Tableau 2. Tableau de scores de résultat diagnostique

4.4. *Évaluation des zones*

Pour chaque zone homogène identifiée, le système doit calculer le taux de certitude ou le "ratio" de chaque couple (voir la formule (1)). Comme ce ratio peut varier de façon importante en fonction du couple, on peut l'utiliser pour l'identification. Il suffit de calculer l'erreur relative entre le ratio du meilleur couple et celui du second couple :

$$q = \frac{l_1 - l_2}{l_1} \qquad (2)$$

avec :

l_1, le taux de certitude maximum calculé selon la formule (1) avec le modèle du couple <langue, codage> correspondant.

l_2, le taux de certitude secondaire calculé avec la formule (1).

L'idée ici est qu'après le calcul d'erreur relative d'une zone, si ce taux est inférieur à une valeur λ, on continue de séparer cette zone en zones plus petites à évaluer.

Exemple : avec λ=0.25 (méthode expérimentale), sur la zone actuelle, nous avons $l_1 = 0.7$, $l_2 = 0.3$, alors $q = (0.7 - 0.3)/0.7 = 0.57$ ($>\lambda$). Nous décidons que cette zone est homogène et choisissons la langue et le codage selon le modèle du couple <langue, codage> du taux l_1. Si $l_1 = 0.7$, $l_2 = 0.6$, alors $q = 0.14$ et nous décidons que cette zone est hétérogène et il faut la séparer en zones plus petites pour l'évaluer ou il faut décider la langue et le codage selon le modèle du couple <langue, codage> de l_1 si cette zone est trop petite pour continuer la séparation.

5. Expérimentation

Nous avons construit un site web pour démonstration d'outil du diagnostic et de la segmentation des textes hétérogènes. Les utilisateurs entrent un texte hétérogène ou choisissent un fichier disponible. Ce système analysera et rendra le résultat qu'il est un couple <langue, codage> utilisé si ce texte est homogène, sinon, il est un texte sous la forme XML avec des balises marquées de couple <langue, codage> d'une zone homogène. Par exemple, avec le texte "*Life is rarely as we would like it to be rather it is exactly as it is : C'est la vie!*", le résultat obtenu est un texte sous la forme XML :

```
<En-CP1252>
    Life is rarely as we would like it to be rather it is     exactly as it is :
</ En-CP1252>
<Fr-CP1252>
    C'est la vie!
</Fr-CP1252>
```

Nous avons testé cet outil avec des textes écrits en plusieurs langues et codages différents comme russe/KOI8, chinois/Big5, japonais/shift-jis...

6. Conclusion et perspective

Nous avons développé un outil en ligne pour analyser des textes homogènes et hétérogènes. Cet outil permet diagnostiquer pour 53 langues et avec plusieurs systèmes de codage différents. De plus, on peut facilement ajouter les autres langues et codages par les fichiers textes d'apprentissage. Cependant, le diagnostic des textes courts n'est pas encore bon. Avec les textes inférieurs à 30 lettres, le taux exact du diagnostic est environ 70% (Vo-Trung, 2003). Nous pensons à utiliser des dictionnaires monolingues pour évaluer des mots et des groupes de mots pour les zones très courts mais il n'est pas encore une bonne solution.

Nous sommes en train d'étudier pour appliquer cet outil dans des systèmes de traduction automatique multilingue. Il sert à segmenter et diagnostiquer la langue pour déterminer les couples langues originales et cibles de traduction de chaque zone monolingue.

Références

Beesley K. R. (1998). Language identifier : A computer program for automatic natural language identification of on-line text. In *Language at Crossroads : Proceedings of the 29th Annual Conference of the American Translators Association*.

Benny G. (2001). *Reconstruction et Utilisation de SILC*. Rapport de stage, Département d'Informatique et de recherche opérationnelle, Université de Montréal.

Bouffard V. (2002). *Évaluation de SILC*. Rapport scientifique, Département d'Informatique et de recherche opérationnelle, Université de Montréal.

Cavnar W. et Trenkle J. (1994). N-gram Based Text categorization. *Symposium On Document Analysis and Information Retrieval*, University of Nevada.

Giguet E. (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multilingues.* Thèse, Université de Caen.

Grefenstette G. (1995). Comparing Two Language Identification Schemes. In *Actes des JADT 1995.*

Manning C. et Schutze H. (5-1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Russell G. (2002). *The QUE Language and Encoding Identification Package*. RALI, Université de Montréal.

Vo-Trung H. (2003). Évaluation des méthodes et des outils actuels pour identifier automatiquement la langue et le codage d'un texte homogène. In *Conférence MAJECSTIC'03*.

Organisation d'une masse documentaire électronique présentée à des lecteurs potentiels

David Wartel, Pascal Francq, Alain Delchambre

CAD/CAM – Université Libre de Bruxelles – 1050 Bruxelles – Belgique
dwartel@ulb.ac.be, pfrancq@ulb.ac.be, adelch@ulb.ac.be

Abstract

Since the early 90's, and thanks to the facilities of communication technologies developed such as the Internet, the number of electronic documents has continuously increased so that it seems more and more difficult and time-consuming for a user, drowned in this documentary space, to find relevant information. To reduce this effect of critical documentary mass and to decrease the access time to relevant resources, we describe in this article a method to organize electronic documents. This organization contains two levels: the first level being described by the concept of "points of view", the second one by the concept of "expert key-words". The method has to be a few time-consuming algorithm in order to be included into a real project.

Résumé

À l'heure où le nombre de documents électroniques ne cesse d'augmenter, et ce depuis plus de dix ans, grâce aux facilités de communication qui se sont développées, notamment l'Internet, il semble de plus en plus difficile et de plus en plus long pour un utilisateur, noyé dans cet espace documentaire, de trouver de l'information pertinente. Pour réduire cet effet de masse documentaire et diminuer le temps d'accès aux ressources importantes, nous décrivons dans cet article une méthode d'organisation de documents électroniques. La méthode doit être très simple et rapide pour s'inclure dans un système informatique réel. L'organisation est à deux niveaux : le premier niveau est décrit par la notion de « points de vue », et le second niveau par la notion de « mots experts ».

Mots-clés : organisation documentaire, information, rapidité d'exécution, « points de vue », « mots experts ».

1. Introduction

Que ce soit pour des recherches dans le domaine du World Wide Web (Berners-Lee *et al.*, 1994), ou au sein des réseaux internes aux entreprises (intranet), la quantité d'information stockée est généralement très importante, et croît quotidiennement, ce qui rend fastidieux l'accès aux informations. Des outils, déjà existants ou en cours de développement, permettent de stocker ces informations de manière à ce que l'utilisateur puisse accéder plus rapidement aux données qui l'intéressent. Parmi eux, le projet GALILEI[1] développe une solution d'analyse et de groupement permettant à des utilisateurs de partager des ressources intéressantes suivant leurs centres d'intérêt.

La méthode que nous décrivons dans cet article s'intègre dans le projet GALILEI. Elle consiste à organiser un ensemble d'éléments (documents électroniques, utilisateurs, ou groupes d'utilisateurs) en un système de mots clés à deux niveaux, que nous appellerons « points de vue » et « mots experts ». Des méthodes déjà existantes, comme les cartes auto organisatrices basées sur les réseaux de neurones (Kohonen, 2000) auraient répondu au problème d'organisation des données, mais la méthode souhaitée doit être rapide en temps de

[1] Subventionné par la région wallonne sous le contrat 01/1/4675 - http://www.galilei.ulb.ac.be

calcul pour pouvoir s'intégrer dans la partie « client » du projet. Elle doit permettre d'alléger la présentation à un utilisateur d'une grande quantité de données, essentiellement des documents, en les organisant sous forme de mots-clés. Il est important de bien noter que le faible temps d'exécution de la méthode est, dans notre cadre, la nécessité la plus importante.

Nous décrirons dans un premier temps le contexte de recherche (section 0), puis les représentations utilisées pour représenter les différents types d'information (section 0), suivront les règles de calcul des « points de vue », des « mots experts » et les méthodes d'organisation (section 0). Nous présenterons enfin la détermination des paramètres et les résultats obtenus sur nos bases de données (section 0).

2. Le projet GALILEI

Le projet GALILEI (Francq, 2003) est un projet utilisant l'indexation de documents électroniques de plusieurs types, permettant le groupement d'utilisateurs et la création de communautés virtuelles. Il a pour objectif d'aboutir à un produit fini, intégrable dans tout système de gestion documentaire. C'est dans le cadre de ce projet que la méthode décrite ici est implémentée, elle doit donc être performante et relativement peu gourmande en temps de calcul.

Pour être reconnu par le système GALILEI, un utilisateur doit se définir comme une personne possédant un ou plusieurs centres d'intérêt. Pour chacun d'entre eux, des profils devront être créés, qui permettront de se connecter au système. La définition d'un profil se fait juste par le choix arbitraire d'un nom que choisir l'utilisateur pour utiliser ce profil. Une fois connecté, le profil peut émettre des jugements sur des documents électroniques locaux, ou des documents accessibles à distance (intra ou inter réseaux). Le jugement des documents se fait de manière rapide et simple : alors qu'il consulte un document, l'utilisateur peut, en un click de souris, définir son document comme intéressant, peu intéressant, ou complètement hors sujet. C'est cette définition que nous utiliserons dans notre méthode pour définir un utilisateur : une personne permettant d'alimenter le système en source d'informations et effectuant un jugement sur certaines d'entre elles.

Les documents jugés sont analysés par le système, ce qui permet de décrire les profils et de les grouper en communautés de même intérêt. Les utilisateurs se verront alors proposer les documents provenant de leurs communautés et qu'ils n'ont pas jugés : les documents les plus intéressants seront donc partagés au sein de chaque communauté. Ces documents peuvent s'avérer être nombreux (plus de 100). Pour éviter d'être équivalente au résultat d'une simple requête dans un moteur de recherche, la méthode présentée va organiser les documents échangés entre utilisateurs et ainsi rendre plus convivial l'accès aux informations pertinentes. Il n'est pas nécessaire que l'ensemble des documents à échanger soit organisé, mais les plus intéressants d'entre eux doivent l'être absolument.

3. Représentations utilisées

L'algorithme d'organisation des documents est un outil qui se greffe au projet GALILEI. Il se base donc sur la représentation utilisée dans le projet : la représentation vectorielle (Salton et McGills, 1983). Cette représentation a principalement été choisie pour sa rapidité au niveau du temps de calcul.

3.1. *Représentation des documents*

Un document électronique est décrit par sa représentation vectorielle dans l'espace des mots. Cet espace est de très grande dimension. Pour réduire l'espace des mots, des traitements addi-

tionnels sont appliqués, comme l'utilisation d'anti-dictionnaires, ou d'algorithmes de désuffixation comme Porter (1980), ou son équivalent français Carry (Paternostre *et al.*, 2002).

Un document est donc perçu comme un vecteur, fortement creux, constitué de poids représentant le nombre d'occurrence du mot dans le document.

$$D_i = \{d_{i,j}\} \quad \text{avec } d_{i,j} \text{ le nombre d'occurrence du mot } j \text{ dans le document } i$$

3.2. *Représentation des utilisateurs*

Les utilisateurs jugent certains documents. A partir des documents jugés comme relevants, nous calculons un vecteur de mots qui représente l'utilisateur. Ces mots représentent les N premiers termes du vecteur somme de tous les documents jugés par l'utilisateur, dont les poids ont été multipliés par un facteur de fréquence inverse (idf) qui représente le logarithme du nombre total de documents divisé par le nombre de documents contenant le mot. Les coordonnées de ce vecteur sont ensuite ordonnées de manière décroissante (les termes de plus fort poids en premier).

$$U_i = \{u_{i,j}\} \quad \text{avec} \quad u_{i,j} = \sum_k \frac{d_{k,j}}{\max_n(d_{k,n})} \times \log\left(\frac{D}{D_j}\right)$$

avec D le nombre total de documents, et D_j le nombre de documents contenant le mot j.

Tous comme les documents, les utilisateurs sont aussi représentés sous forme de. Remarquons cependant que, contrairement aux documents, les utilisateurs bénéficient du facteur idf. Celui-ci est nécessaire pour le système GALILEI afin d'améliorer la création de groupes. La méthode, en soit, reste la même : les utilisateurs sont identifiés par un vecteurs de mots ordonné par occurrences décroissantes.

3.3. *Représentation des groupes*

Les utilisateurs étant définis, ils sont regroupés en communautés virtuelles grâce à des algorithmes de groupement comme le k-Means (MacQueen, 1967 ; Wartel 2002) ou le *Genetic Virtual Community Algorithm* (Francq, 2003) Ces méthodes de groupements utilisent entre autre la similarité, définie comme étant le cosinus entre deux vecteurs mots. Les utilisateurs partageant les mêmes centres d'intérêt sont alors groupés ensemble. Là encore, nous calculons une représentation vectorielle des groupes ainsi obtenus.

4. Organisation

Les documents, les utilisateurs et les groupes sont donc représentés par des vecteurs de mots dans l'espace des mots. La méthode d'organisation a pour but de cartographier ces vecteurs en deux niveaux. Que les vecteurs soient des documents, des groupes ou des utilisateurs, la méthode d'organisation reste la même. La méthode est basée sur une analyse statistique de l'occurrence des mots au sein des données à organiser. Nous n'allons retenir des vecteurs d'entrée que les mots de plus forte occurrence. Intuitivement, on ressent que ces mots représentent les éléments les plus importants, nous allons donc les utiliser pour définir des « points de vue » et des « mots experts » qui permettront d'organiser l'ensemble des vecteurs.

4.1. *Paramètres*

La méthode d'organisation requiert certains paramètres de configuration :

- ***N***, représentant le nombre de mots clés à calculer.
- ***P***, représentant le pourcentage de « points de vue » désiré.

4.2. *Calcul des « points de vue » – « mots experts »*

Les « points de vue » correspondent aux *N* mots les plus importants (ou les plus « représentatifs » sur l'ensemble des vecteurs d'entrée *Ve*. Rappelons que ces vecteurs d'entrées peuvent être l'ensemble des groupes utilisateurs, l'ensemble des utilisateurs d'un groupes, ou l'ensemble des documents d'un utilisateur.

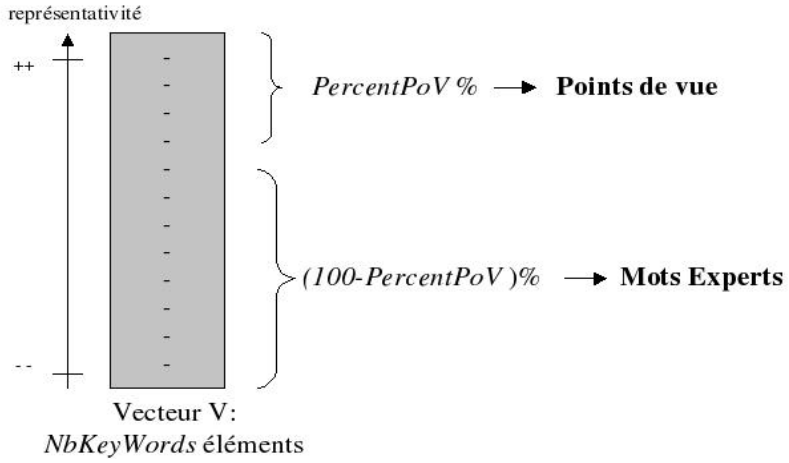


Figure 1. Séparation des «points de vue» et «mots experts»

Comme pour le calcul des utilisateurs, nous allons identifier les mots les plus importants de tous les vecteurs en calculant le vecteur *V* somme des vecteurs d'entrée, chaque terme étant multiplié par son facteur idf :

$$V = \{v_i\} \qquad \text{où} \qquad v_i = \sum Ve_i \times \log\left(\frac{N_v}{Nv_i}\right)$$

avec N_v le nombre total de documents et N_{vi} le nombre de documents contenant le mot i.

Nous obtiendrons donc *N* mots.

Parmi ceux-ci, les *P* premiers pourcents seront considérés comme « points de vue » pour nos vecteurs d'entrée. Les (1-P) autres pourcents seront leurs « mots experts » (*Figure 1*).

4.3. *Organisation des « mots experts »*

Pour obtenir notre organisation à deux niveaux, il faut associer les « mots experts » aux « points de vue ». Pour cela, nous allons calculer pour chaque « mot expert » le « point de vue » avec lequel il a la cooccurrence la plus forte sur l'ensemble des vecteurs d'entrée. La cooccurrence d'un mot est définie comme le nombre de fois que les mots apparaissent ensemble sur l'ensemble des vecteurs d'entrée *V*:

$$Co(m_1, m_2) = card(v_1 \cap v_2)$$

avec v_1 l'ensemble des vecteurs contenant le mot *m1* et v_2 l'ensemble des vecteurs contenant le mot *m2* :

$$v_1 = \{V_1 \subset V \mid \forall \vec{v} \in V_1, \vec{v}.\overrightarrow{m_1} \neq 0\}$$

$$v_2 = \{V_2 \subset V \mid \forall \vec{v} \in V_2, \vec{v}.\overrightarrow{m_2} \neq 0\}$$

Un « mot expert » ne pourra donc être associé qu'à un et un seul « point de vue » (*Figure 2*).

Il peut arriver qu'un « mot expert » ne trouve pas de « point de vue » lui correspondant : cela voudra dire que ce « mot expert » n'apparaît avec aucun « point de vue » sur l'ensemble des vecteurs d'entrée. Il est important de quantifier cette « perte » de « mots experts », et de vérifier que celle ci reste faible. Au cas où celle-ci devient trop forte, les paramètres *N* et *P* devront être réajustés. De même, il arrive qu'un « point de vue » reste vide, c'est-à-dire qu'aucun « mot expert » ne lui est attribué. Ce cas est moins grave : le « point de vue » peut être supprimé si l'organisation des données le permet.

```
M={ensemble des Mots Experts}
Tant que M non vide
    m=premier mot de M
    P={ensemble des Points de vue}
    best_occurrence = 0
    best_pov = 0
    Tant que P non vide
        p=element pris au hasard dans P
        Calculer C=co_occurrence(m,p)
        Si C > best_occurrence
            best_occurrence=C
            best_pov=p
        Supprimer p de P
    Si best_pov different de 0
        affecter m à best_pov

    Supprimer m de M
```

Figure 2. Algorithme d'organisation des «mots experts»

4.4. *Organisation des données*

La hiérarchie à deux niveaux réalisée, les vecteurs d'entrée, qui correspondent aux données à catégoriser, peuvent être classés suivant les « mots experts ». L'organisation des données est simple : un vecteur d'entrée est affecté à un « mot expert » si et seulement si le «mot expert» apparaît dans ce vecteur et sa fréquence est la plus importante de toutes les fréquences des « mots experts » pour ce vecteur d'entrée.

Ainsi, soit $\vec{m}_e$ un « mot expert », $\vec{v}$ un vecteur d'entrée :

$$\vec{v} \in \vec{m}_e \quad \Leftrightarrow \quad \vec{v}.\vec{m}_e \neq 0 \text{ et } \forall \vec{m}_i \in \{mots \text{ exp } erts\}, \vec{m}_e = \arg\min_i (\vec{v}.\vec{m}_i)$$

Nous avons choisi dans notre cas de n'affecter un même vecteur d'entrée qu'à un et un seul « mot expert ». Il suffit juste que ce mot apparaisse dans le document. Si on veut qu'un vecteur puisse appartenir à plusieurs mots experts différents, on pourra juste vérifier que le mot expert appartient à ce vecteur :

$$\vec{v} \text{ affecté } \vec{m}_e \quad \Leftrightarrow \quad \vec{v}.\vec{m}_e \neq 0$$

De même que pour l'organisation des « mots experts », il est ici aussi important de mesurer la perte de vecteurs d'entrée, c'est-à-dire le nombre de vecteurs d'entrée non affectés à au moins un « mot expert ». Si ce nombre excède une valeur seuil, les paramètres *N* et *P* devront être

modifiés. La valeur seuil pourra être le nombre moyen de documents affectés à un même « mot expert ».

5. Dimensionnement : application sur un ensemble de groupe de nouvelles

Nous avons vu que l'algorithme dépend principalement de deux paramètres : *N* et *P*. Rappelons que l'objectif est d'obtenir une organisation rapide, sans trop de perte de données et minimisant le nombre de « points de vue » et de « mots experts » perdus (i.e., sans élément affecté).

L'algorithme d'organisation a été testé sur deux bases de données. La première (appelée news1500), issue de groupes de nouvelles (newsgroups) contient environ 1500 documents réunissant 76 utilisateurs en 18 groupes. La seconde (appelée news20000), toujours issue de groupes de nouvelles, contient environ 20000 documents, et regroupe 140 utilisateurs en 18 groupes. L'algorithme a été écrit en C++, et les résultats traités et analysés avec le logiciel R. Nous nous attachons, dans ce dimensionnement, à l'organisation des documents pour chaque groupe.

Pour les tests, le nombre de mots clés *N* a évolué de 10 à 100 pour le classement des documents. En effet, au delà de cette valeur, nous considérons que le nombre de « points de vue » devient trop important pour l'utilisateur. Le pourcentage de « points de vue » *P varie* de 10 % à 50 %. Au delà, nous aurions forcé certains « points de vue » à être vides.

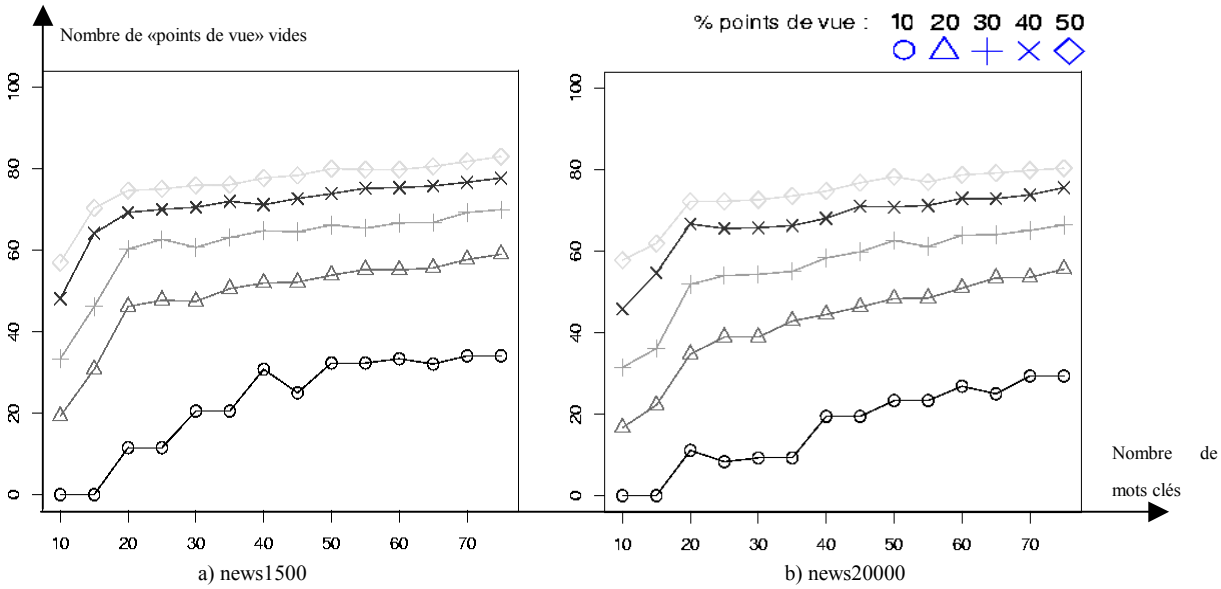


Figure 3. Nombre de «points de vue» vides pour différents pourcentages de «points de vue»

Si le nombre *N* devient trop grand, on peut arriver à des pourcentages élevés de « points de vue » vides, aussi bien sur la base de données news1500 que la base de données news20000.

Cet effet est légèrement amplifié si on augmente le pourcentage de « points de vue » parmi le nombre total de mots clés (*Figure 3.*). Ce constat est relativement normal : plus N est grand, plus nous allons considérer des mots peu importants comme mots clefs, et donc comme points de vue. Ces mots n'ont pas de cooccurrence assez élevée avec les mots experts, ils restent alors vides.

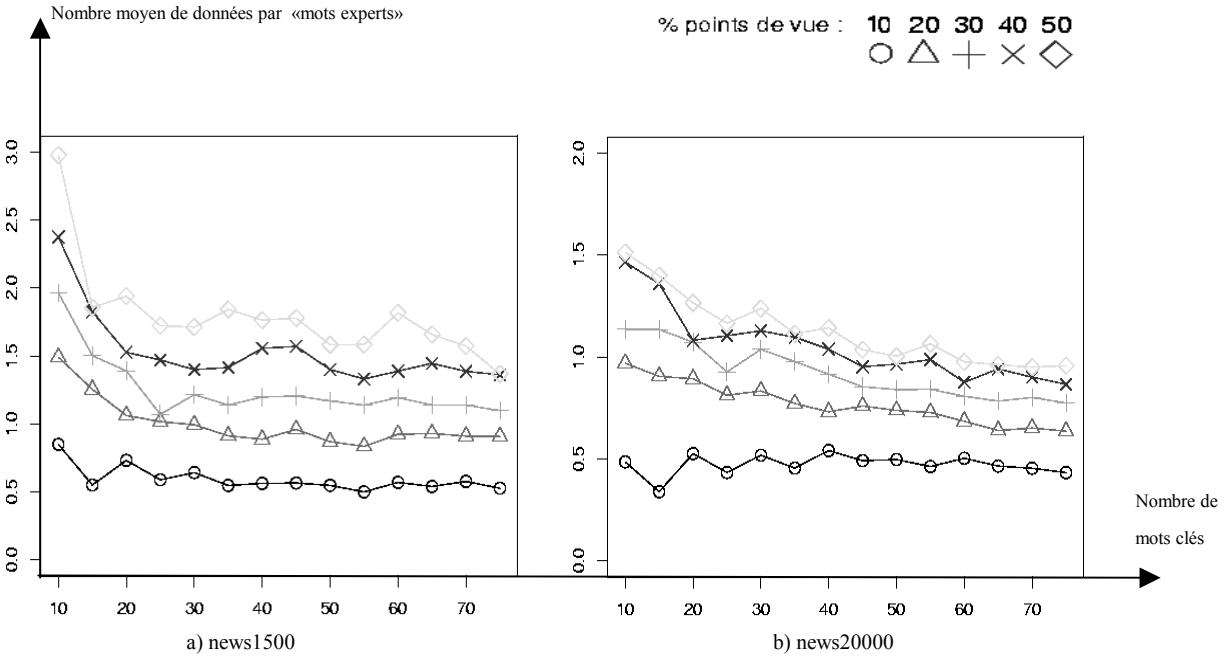


Figure 4. Nombre moyen de données par «mots experts» pour différents pourcentages de «points de vue»

Le fait que certains « points de vue » soient vides n'est, en soi, pas problématique : on peut les supprimer mais il faut cependant vérifier que le nombre de « points de vue » restants soit acceptable, et que le nombre de « mots experts » non affectés à un « point de vue » et le nombre de données non classées soient peu élevés. Dans notre cas, le nombre de « mots experts » non affectés est nul pour les deux bases de données: chaque « mot expert » a trouvé un « point de vue » lui correspondant.

De plus, on pourrait penser que le fait d'augmenter le nombre de mots clés conduirait à une organisation plus rigoureuse. Cependant, le nombre moyen de documents par « mots experts » varie peu (*Figure 4.*). En effet, l'augmentation du nombre de mots clés s'accompagne d'une augmentation du nombre de « mots experts » vides (*Figure 5.*). L'explication est la même que pour les « points de vue » vides : en augmentant le nombre de mots clefs, et donc de « mots experts », on retient des mots dont la fréquence d'apparition dans les documents n'est pas assez importante. Aucun document ne leur est donc affecté, ils restent vides.

Le nombre de données non affectées à des « mots experts » est également un paramètre important. Il correspond au nombre de données que le système n'indexera pas, elles seront donc considérées comme inexistantes ou perdues. Le nombre de ces données diminue avec l'augmentation du nombre de mots clés (*Figure 6.*). Si ce nombre reste trop faible (20), nous pouvons obtenir entre 20 et 40 % de perte pour la base de données news20000. Ces pertes deviennent inférieures à 10 % pour un nombre moyen de 50 mots clés. Il faut donc, pour cette base de données, utiliser un nombre assez important de mots clés pour minimiser cette perte,

le peu de documents perdus pourront être classés sous un autre mot expert crée pour regrouper les données non classées.

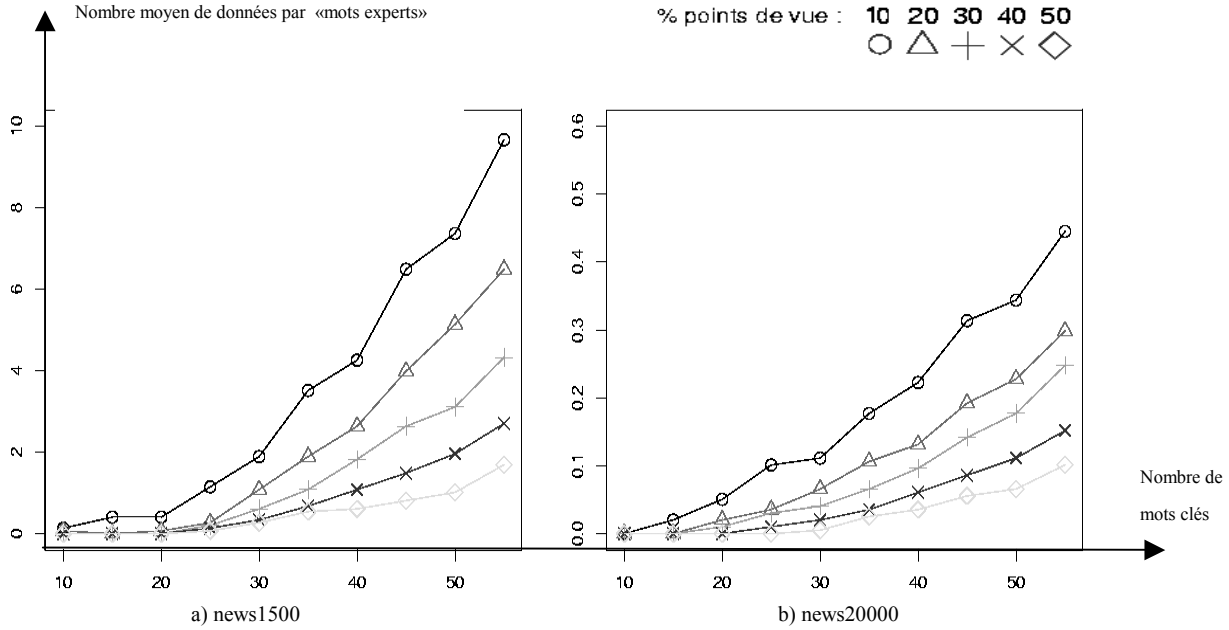


Figure 5. Nombre de « mots experts » vides pour différents pourcentages de « points de vue »

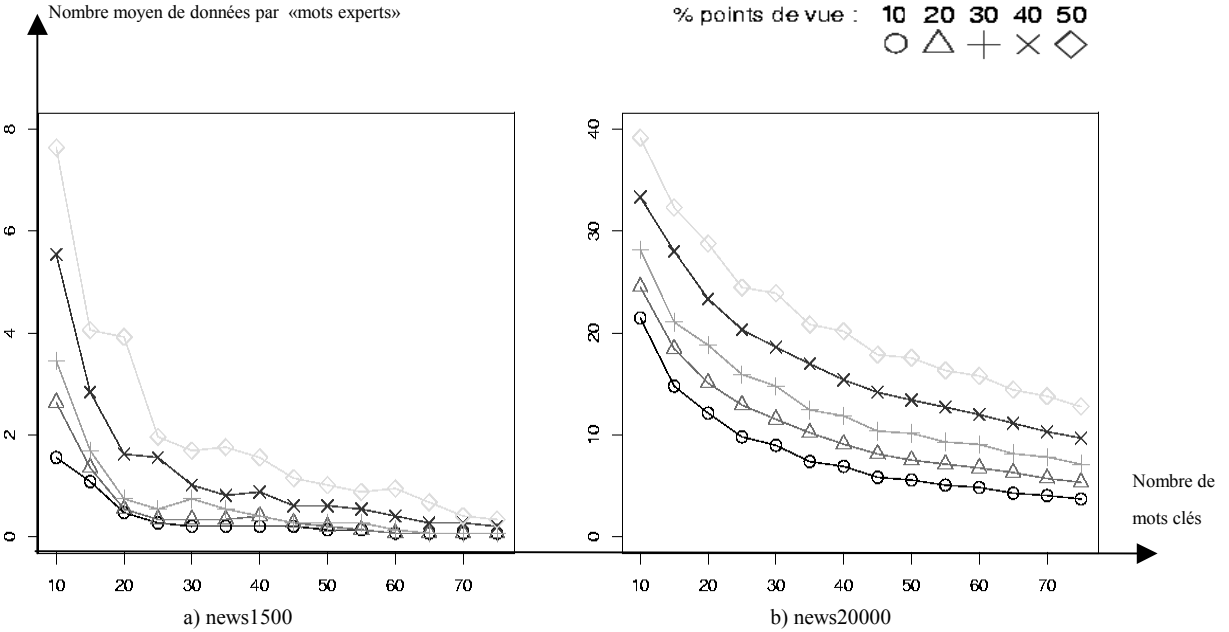


Figure 6. Nombre de documents non classés pour différents pourcentages de « points de vue »

Précisons cependant que ces documents non classés sont en grande partie les documents les moins intéressants à l'intérieur des communautés d'utilisateurs. En effet, ceux-ci sont perdus car ils ne contiennent aucun « mot expert », qui sont les mots importants sur l'ensemble des documents. Nous pouvons donc penser que les documents perdus sont des documents qui

n'apportent pas l'information pertinente à l'intérieur du groupe. Leur perte n'est donc pas un problème majeur.

Le pourcentage de « points de vue » doit rester, sur l'ensemble des résultats, relativement faible. Ainsi, il permet de minimiser le nombre de « points de vue » vides (*Figure 3.*), et d'obtenir un nombre faible de documents non affectés, et donc peu de perte de données (*Figure 6.*). Dans notre cas, un pourcentage de « point de vue » de 20 permet d'obtenir de bons résultats.

Dans cette configuration, nous obtenons des résultats très rapidement ce qui satisfait aux conditions d'intégrations dans notre système.

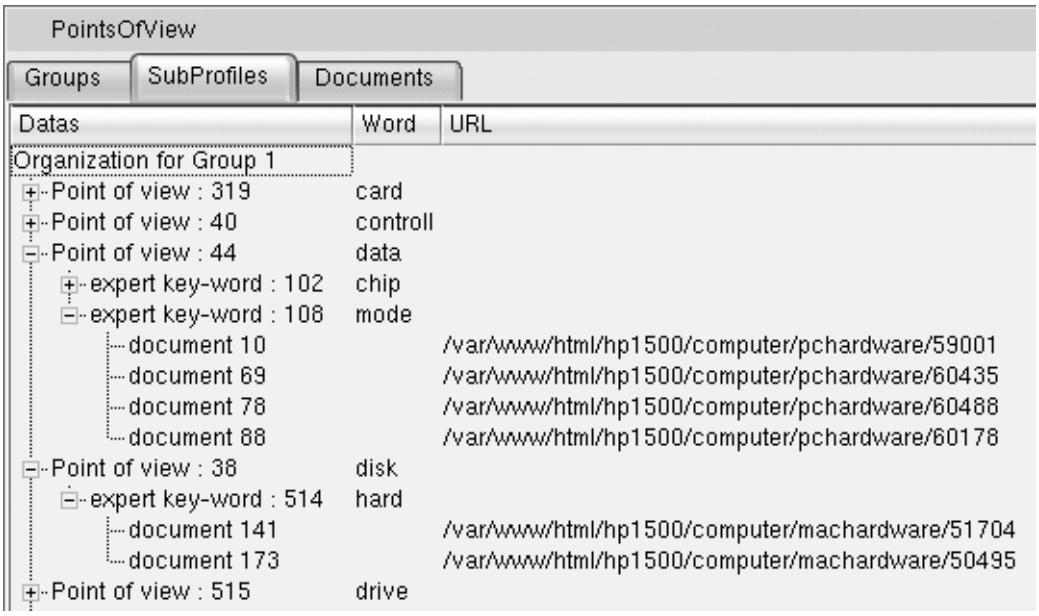


Figure 7. Présentation des points de vue dans GALILEI

La *Figure 7.* montre le résultat du découpage en points de vue et mots experts de l'ensemble des documents jugés par un utilisateur au sein du système GALILEI. L'intérêt global de l'utilisateur est l'informatique. Ainsi, nous remarquons une organisation en points de vue tels que « card », « controller », « data » ou encore « disk », et un sous découpage, pour chacun de ces points de vue, en mots experts tels que « chip », « mode » ou « hard ». L'organisation n'est certainement pas parfaite mais là n'est pas l'objectif recherché : les documents sont organisés de manière à adoucir l'effet de la masse documentaire ce qui permet une navigation plus conviviale pour l'utilisateur.

6. Conclusion

La méthode présentée dans cet article permet une organisation paramétrable d'un ensemble de documents. Cette organisation se fait à deux niveaux, un niveau de « points de vue » et un niveau de « mots experts », et, grâce à la même représentation vectorielle utilisée, elle peut s'appliquer sur trois types d'entités informationnelles : des groupes, des utilisateurs, et des documents. Pour chaque type d'entité, une étape de dimensionnement permet de définir les paramètres à utiliser pour l'organisation des données. Actuellement, des études sont menées pour trouver un dimensionnement automatique. Dans le cas étudié, il a été montré que pour garder une organisation significative et ne pas obtenir un nombre trop important de données

perdues, il nous faut un nombre de mots clés relativement faible (20) pour la base de données news1500, et moyennement élevé (45) pour la base de données news2000. Ces configurations conduisent tout de même à des pertes de documents, pertes qui ne sont pas très importantes vue qu'elles correspondent à des documents peu pertinents. Par ailleurs, le pourcentage de « points de vue » parmi ces mots clés doit lui rester faible (20 %), pour éviter d'accentuer certains effets néfastes du nombre de mots clés comme le nombre de points de vue vides ou le nombre de données perdues.

L'organisation des documents est rapide (moins de 3 secondes pour une base de données de 1500 documents et 18 groupes) et permet donc une implémentation dans le projet GALILEI.

Références

Berners-Lee T., Caillau R., Luotonen A., Nielsen H. et Secret A. (1994). The world-wide web. In *Communication of the ACM* : 76-82.

Francq P. (2003). *Structured and Collaborative Search: An integrated approach to share documents among users*. PhD Thesis, Université Libre de Bruxelles.

Kohonen T. (2000). *Self-Organizing Maps* (3rd edition). Springer Verlag.

MacQueen J.(1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability*, vol.(1) : 281-297.

Paternostre M., Francq P., Saerens M., Lamoral J. et Wartel D. (2002). *Carry, un algorithme de désuffixation pour le français*, URL : http://www.galilei.ulb.ac.be

Porter M.F. (1980). An Algorithm for suffix stripping. *Program*, vol. (14) : 130-137.

Salton G. et McGills M. (1983). *Modern Information Retrieval*, McGraw-Hill Book Co.

Wartel D.(2003). *Les algorithmes de clustering dans GALILEI,* rapport scientifique, Université Libre de Bruxelles.

Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles

Maria Zimina

EA 2290 SYLED, Université de la Sorbonne nouvelle – Paris 3 – 19, rue des Bernardins – 75005 Paris – France – zimina@msh-paris.fr

Abstract

The approach suggested in this article enables statistic identification of multiple word correspondences in bilingual texts aligned on phrase level. Corresponding lexical units are discovered through characteristic element computation in parallel contexts. The calculation undergoes systematic reiteration within a corpus. The exploratory results show that the use of quantitative methods in combination with bilingual text topography offers new means for automatic description of lexical equivalences.

Résumé

Dans les corpus bilingues alignés au niveau de la phrase, le repérage de correspondances lexicales multiples peut être effectué sur des bases quantitatives. Fondée sur la pratique du calcul des spécificités, notre méthode explore parallèlement les contextes équivalents pour repérer des correspondances dans les emplois caractéristiques des différents types d'unités textuelles. La réitération systématique de ce processus dans le corpus, appuyée sur l'utilisation des principes de topographie bi-textuelle, offre de nouveaux moyens automatisés pour l'extraction et la description d'équivalences lexicales.

Mots-clés : corpus parallèles, alignement lexical, correspondances de traduction, textométrie, topographie bi-textuelle.

1. Les corpus parallèles alignés

Dans le contexte multilingue, les corpus de documents parallèles sont constitués de plusieurs volets qui correspondent chacun à une version d'un même texte dans plusieurs langues différentes. Il s'agit, en général, de textes sources et de leurs traductions (effectuées par des traducteurs humains) présentés sous forme électronique ou de textes dont chacun est une traduction de l'autre sans qu'il soit possible de déterminer lequel a servi de source. Le traitement de corpus parallèles suppose une phase préalable d'*alignement*, c'est-à-dire la mise en correspondance d'unités textuelles de différents types entre chacun des volets du *bi-texte*.

De nombreux travaux ont montré l'utilité des corpus parallèles alignés pour le développement des applications de traitement automatique des langues (Isabelle et Warwick-Armstrong, 1993 ; Véronis, 2000). La création automatique des lexiques bilingues alignés permet de normaliser la terminologie dans un champ de communication. Elle est indispensable à la maintenance de la documentation technique multilingue. Les travaux sur les corpus alignés contribuent aussi au développement des moteurs de recherche multilingues sur le Web. Pour le traducteur, les corpus alignés fournissent un accès au savoir-faire de la communauté lorsqu'il s'agit de traduire une expression pour laquelle les solutions proposées par des ouvrages de référence (dictionnaires, bases de données terminologiques, etc.) sont insatisfaisantes. En lexicographie, l'alignement des données de traduction met en lumière la spécificité et la

richesse du vocabulaire bilingue. Les corpus alignés permettent de découvrir des usages et des expressions ne figurant pas encore dans les dictionnaires.

2. Alignement lexical : problèmes et enjeux

Les comptes-rendus d'expériences publiés récemment décrivent des algorithmes permettant d'apparier les phrases d'un corpus parallèle avec un taux de réussite élevé (Véronis, 2000)[1]. En revanche, l'*alignement lexical*, c'est-à-dire la mise en correspondance de mots et locutions entre les deux volets d'un corpus parallèle demeure un problème difficile. Lors de l'alignement de ce type d'unités, il faut tenir compte de plusieurs phénomènes complexes liés, notamment, à la détection des emplois polysémiques de mots et de leur fonctionnement dans des séquences figées et locutions dont la traduction varie selon le contexte.

3. Corpus *Convention* : navigation textométrique

3.1. *Unités lexicales à correspondances multiples*

Lorsqu'il s'agit de mots dotés d'un large éventail de sens, les correspondances traductionnelles entre les deux volets d'un corpus parallèle forment un réseau complexe. Les unités lexicales d'un volet peuvent recevoir plusieurs traductions dans l'autre volet. Pour illustrer ce type d'équivalences, nous emprunterons des exemples à un corpus de textes juridiques anglais-français de la *Convention de sauvegarde des droits de l'homme et des libertés fondamentales*, désormais ***Convention*** (cf. *Tableau 1*)[2].

volet français	volet anglais
conv_a0_p1-1 **1** Les gouvernements signataires , membres du Conseil de l'Europe,	**conv_a0_p1-1e** **2** The governments signatory hereto, being members of the Council of Europe,
conv_a0_p2-1 **3** Considérant la Déclaration universelle des Droits de l'Homme, proclamée par l'Assemblée générale des Nations Unies le 10 décembre 1948 ; /…/	**conv_a0_p2-1e** **4** Considering the Universal Declaration of Human Rights proclaimed by the General Assembly of the United Nations on 10th December 1948; /…/
conv_a0_p7-1 **13** Sont convenus de ce qui suit /…/	**conv_a0_p7-1e** **14** Have agreed as follows /…/
Guide de lecture : Chaque couple de phrases équivalentes est introduit par un code. Les numéros indiquent (dans l'ordre) : le type de document (convention, protocole, etc.) ; le numéro de l'arrêt ; la partie de l'arrêt ; le numéro de la section et/ou du paragraphe ; le numéro de la phrase dans le corpus, précédé par la lettre « e » pour les phrases en anglais.	

Tableau 1. Corpus Convention *(extrait)*

Dans le corpus ***Convention***, le mot français *mort* (F=44) est le plus souvent traduit en anglais par *death* (F=26). On rencontre cette correspondance au sein des équivalences lexicales *la*

[1] Les systèmes actuels d'alignement des phrases de textes parallèles multilingues ont fait récemment l'objet d'une étude d'évaluation menée au sein du projet ARCADE. Les résultats de l'étude témoignent d'avancées méthodologiques importantes dans les techniques d'alignement des phrases. Lorsque les textes ne présentent pas de divergences importantes au niveau structurel (pas d'omissions, etc.), le taux de précision des systèmes évalués est estimé, en moyenne, à 98,5 % (cf. Véronis, 2000).

[2] Le corpus a été constitué à partir des documents contenus dans la Convention, d'une douzaine de protocoles, et d'une série d'arrêts rendus par la Cour européenne des Droits de l'Homme de Strasbourg en 1995. Chaque volet du corpus compte approximativement 300 000 mots graphiques. On peut trouver les textes de la Convention en anglais et en français sur le site officiel de la Cour Européenne des Droits de l'Homme : http://www.echr.coe.int.

peine de mort – death penalty, *les circonstances de la mort – the circumstances of the death*, *un danger de mort – a risk of death* et beaucoup d'autres. Cependant, l'équivalence *mort – death* n'est pas préservée dans les contextes suivants :

volet français	volet anglais
La mort n'est pas considérée comme infligée en violation de cet article /…/.	**Deprivation of life** shall not be regarded as inflicted in contravention of this article /…/
/…/ l'article vise certes les cas où **la mort** a été infligée intentionnellement, mais ce n'est pas son unique objet.	/…/ this provision extends to, but is not concerned exclusively with, intentional **killing**.

(Corpus ***Convention***)

Dans les deux couples de phrases ci-dessus le mot *mort* est utilisé dans le sens « fin provoquée de la vie », répertorié, par exemple, par le dictionnaire *Le Robert* (édition 1996). L'absence de cette signification dans l'univers sémantique du mot anglais *death* explique la présence des équivalences *mort – deprivation of life*, *mort – killing* dans le corpus. Comme on va le voir ci-dessous, une description automatique de correspondances traductionnelles multiples peut être envisagée si l'on fait appel à des méthodes d'analyse quantitative.

3.2. *Topographie bi-textuelle et outils statistiques d'appariement*

Les travaux récents montrent qu'il est possible de faire appel à des principes d'*analyse textométrique*[3] et, notamment, à une *représentation topographique*[4] du texte pour l'appariement des mots et des syntagmes des corpus parallèles (cf. Lamalle et Salem, 2002 ; Zimina, 2000 et 2002). La cartographie des *présence-absence* de correspondances bilingues au sein des traductions fournit des moyens automatisés pour le recensement des équivalences. Dans ce qui suit, nous allons décrire une méthode d'analyse permettant la découverte de correspondances lexicales entre les deux volets parallèles pour des mots qui possèdent plusieurs traductions au sein d'un même corpus. Une approche hybride qui allie la *topographie textuelle* et l'*analyse des spécificités*[5] est à l'origine de cette méthode. La recherche de correspondances lexicales s'appuie sur l'alignement du corpus au niveau de la phrase. L'exploration débute par le marquage au fil du texte dans l'un des volets du corpus bilingue d'un sous-ensemble d'occurrences correspondant à un *type* quelconque (*forme graphique*, *lemme*, *segment répété*, etc.). Le repérage des phrases correspondantes dans l'autre volet permet de construire deux fragments de texte équivalents qui seront confrontées dans chacun des cas au reste du corpus à des fins de comparaison.

[3] Sur la description de pratiques de la textométrie, on consultera Heiden (2002).

[4] La *topographie textuelle* a pour objectif une localisation graphique des phénomènes mis en évidence par l'étude statistique.

[5] Le repérage des *spécificités* ou vocabulaires caractéristiques met en évidence, pour un groupe de phrases donné, les unités dont la fréquence connaît une variation importante dans ce fragment de texte. Fondée sur le *model hypergéométrique*, la méthode des spécificités permet d'effectuer une comparaison entre l'ensemble du corpus (*T*) et l'échantillon des contextes contenant l'unité pôle (*t*). En fonction de la fréquence globale des unités attestées dans ce fragment (*F*) et de leur fréquence locale (*f*), on leur affecte un indice de spécificité. Le diagnostic est fourni sous la forme *±Exx* où le signe indique un *sur-emploi* ou un *sous-emploi* de l'unité et la valeur indique son degré de spécificité, cf. (Lafon, 1984 ; Lebart et Salem, 1994).

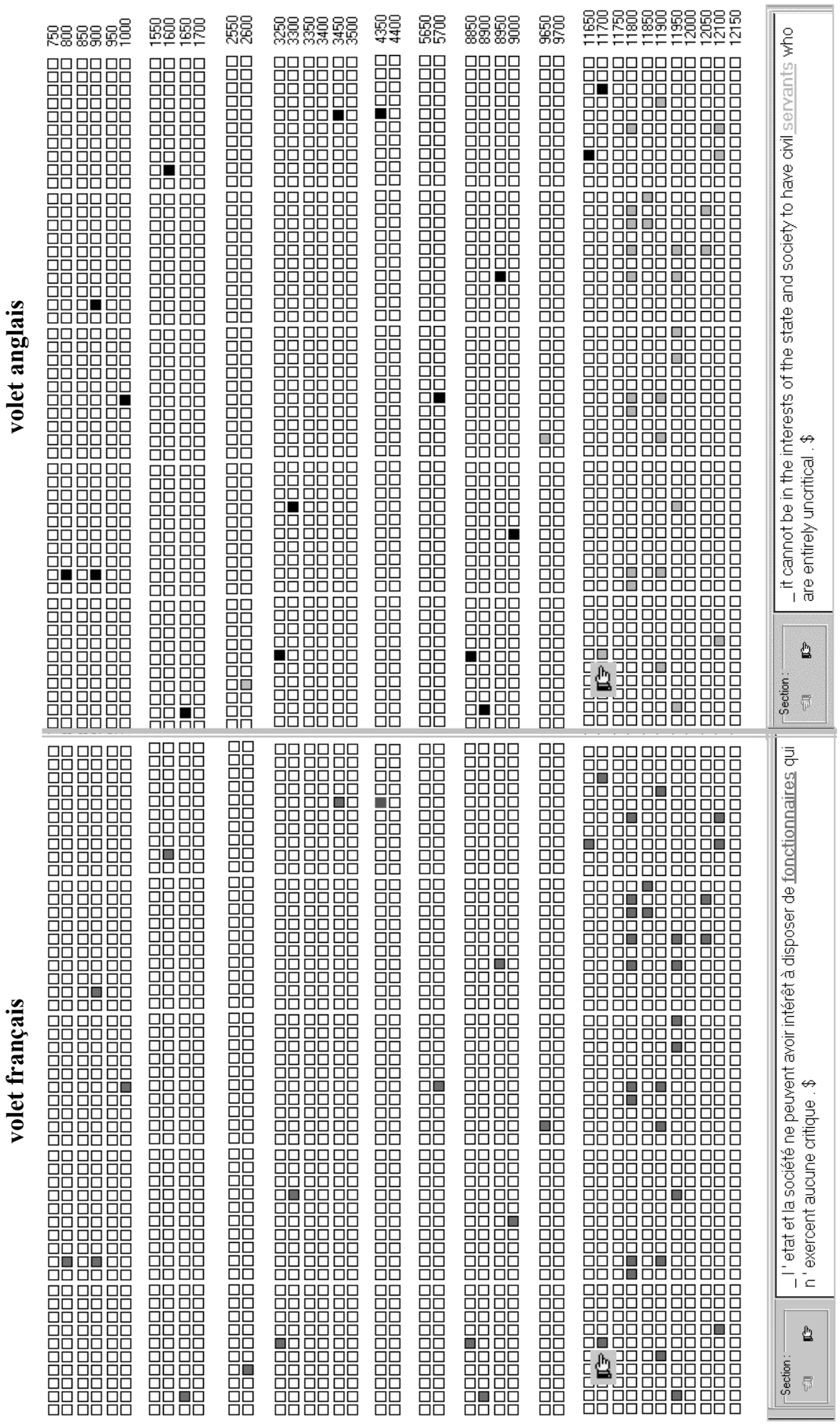


Figure 2 : Un extrait de la carte des phrases *appariées issues du corpus* Convention

Guide de lecture de la figure 2 : Les fonctionnalités développées au sein du logiciel ***Lexico3*** permettent d'envisager la visualisation de l'alignement des phrases d'un corpus parallèle à l'aide d'une *carte des sections*. Sur la figure 2, l'appariement des phrases d'un extrait du corpus ***Convention*** est matérialisé par des carrés positionnés sur deux colonnes. La ventilation de la forme *fonctionnaires* est représentée par les carrés gris foncé ■. Dans le fragment équivalent en anglais, les carrés gris clair ■ correspondent aux occurrences de l'une de ces traductions en anglais, la forme *servants*. La sélection des phrases activée sur la carte est matérialisée par des carrés noirs ■. Elle correspond à des phrases du volet anglais où la forme *fonctionnaires* reçoit d'autres traductions que *servants*.

Note : La description des fonctionnalités de la *carte des sections* prend en compte l'état final d'un ensemble de procédures en cours de développement qui devraient être disponibles dans la prochaine version de ***Lexico3***.

3.2.1. *Construction de la carte des sections*

Les fonctionnalités développées au sein du logiciel ***Lexico3*** (Lamalle *et al.*, 2003) permettent à l'utilisateur de visualiser une *carte des sections*[6] (ex.: phrases ou paragraphes du corpus), puis de constituer une sélection arbitraire de sections dont on étudiera ensuite le vocabulaire spécifique. L'utilisateur dispose d'un ensemble d'outils permettant de choisir (à partir du dictionnaire, du *garde-mots*[7], de la liste des *segments répétés*, etc.) un *type* sur lequel portera son exploration. Après avoir sélectionné le type, il est possible de le faire glisser sur la carte (*glisser/déposer*). La ventilation du type étudié devient alors visible. Les sections dans lesquelles il est présent sont coloriées[8]. Ce processus peut être réitéré.

Pour illustrer notre propos, nous considérerons les phrases du corpus ***Convention*** qui contiennent la forme *fonctionnaires* (F=49) et le sous-ensemble de phrases équivalentes en anglais. Sur la figure 2, l'appariement des phrases du corpus est matérialisé par des carrés positionnés sur deux colonnes. Les carrés sont liés et toute sélection dans un volet est automatiquement répercutée dans l'autre. La ventilation de la forme française *fonctionnaires* est représentée par des carrés gris foncé. L'appariement des phrases étant représenté sur la carte, il est possible d'envisager l'activation de la sélection parallèle des phrases équivalentes dans le volet anglais[9].

3.2.2. *Les outils textométriques de la carte des sections*

Le calcul des *spécificités* permet de sélectionner parallèlement pour chacun des fragments

[6] La ***carte des sections*** permet une visualisation du corpus découpé en sections par la promotion d'un (ou de plusieurs) caractères particuliers au statut de délimiteurs de section (Lamalle *et al.*, 2003). Dans le cas des corpus parallèles, le découpage en sections peut être effectué parallèlement, en s'appuyant sur des codes attribués aux phrases en correspondance.

[7] Le *garde-mots* est une fonctionnalité de ***Lexico3*** permettant de mémoriser formes, segments, etc. pour une utilisation ultérieure. Pour stocker un type dans le *garde-mots* il suffit de le faire glisser sur l'icône de cette fonctionnalité.

[8] Il est possible d'obtenir un coloriage plus ou moins sombre des sections en cochant d'abord la case *seuil*. Cette fonctionnalité permet de régler deux seuils en probabilités qui entraîneront l'intensité du coloriage sur la carte.

[9] La réalisation informatique de la sélection parallèle des sections correspondantes est actuellement en cours de développement. Elle s'appuie sur les fonctionnalités existantes de la *carte des sections* de ***Lexico3*** et utilise des principes formels de la segmentation parallèle des corpus bilingues où chaque couple de phrases équivalentes est introduit par le même code.

ainsi constitués une série d'unités textuelles caractéristiques de ces fragments[10]. Les listes des spécificités s'affichent dans les fenêtres des deux côtés de la *carte des sections bi-textuelle*. Le *Tableau 3* présente quelques-unes de ces unités spécifiques mises en évidence par cette méthode. On constate sur ce tableau que les unités bilingues issues de l'exploration constituent bien des correspondances de traduction. Ainsi, la forme française *fonctionnaires* (*spec*.+E109), ayant servi de point d'entrée pour la construction de l'échantillon de phrases pour cette analyse, peut être appariée avec la forme *servants* (*spec*.+E55) et le segment répété *civil servants* (*spec*.+E52) qui se sont révélés les plus caractéristiques du fragment anglais (cf. *Tableau 3*).

forme / segment	**Frq. Tot.**	**frq. locale**	**Spec.**		**forme / segment**	**Frq. Tot.**	**frq. locale**	**Spec.**
fonctionnaires	49	49	+E109		servants	50	31	+E55
les fonctionnaires	14	14	+E31		civil servants	46	29	+E52
des fonctionnaires	14	14	+E31		civil	304	41	+E40
de loyauté	36	14	+E22		loyalty	43	14	+E20
loyauté	42	14	+E21		duty	109	15	+E16
de loyauté politique	22	10	+E17		political loyalty	25	10	+E16
loyauté politique	24	10	+E16		of political	29	10	+E15
de fonctionnaires	7	7	+E16		duty of	45	11	+E15
obligation de loyauté politique	15	8	+E14		officers	38	10	+E14
obligation de loyauté	21	8	+E13		duty of political loyalty	23	9	+E14
					service	110	14	+E14
					civil service	58	11	+E13
					CONSTITUTIONAL	146	14	+E13

Guide de lecture : Le tableau représente un extrait de la liste des spécificités positives majeures calculées pour l'échantillon des phrases où est attestée la forme *fonctionnaires* et le fragment correspondant en anglais.

Tableau 3. Spécificités majeures pour la sélection bi-textuelle

La fréquence globale de la forme *fonctionnaires* (F=49) est supérieure à celle de la forme anglaise *servants* dans le fragment (f_{locale}=31). Nous pouvons en conclure que la forme *fonctionnaires* reçoit d'autres traductions dans le corpus. Pour découvrir l'ensemble des équivalences lexicales correspondant à la forme-pôle, on soumet au calcul des spécificités les seules phrases du fragment anglais dans lesquelles la forme *servants* est absente (cf. *Figure 2*). La réitération du calcul des spécificités dans ce sous-ensemble de phrases met en évidence une série d'unités les plus caractéristiques de ce nouveau fragment :

[10] Actuellement, les outils statistiques de la *carte des sections* rendent possible une sélection automatique des sections dans lesquelles le type étudié est présent (c'est cet ensemble de sections que l'on compare à l'ensemble du corpus).

forme / segment	Frq. Tot.	frq. locale	Spec.
officers	38	10	+E19
officials	16	7	+E16
senior	18	6	+E13
senior police	5	4	+E11
police	216	9	+E10
senior police officers	3	3	+E09

Le retour au contexte confirme que les unités *officers* (*spec*.+E19), *senior police officers* (*spec*.+E09), *officials* (*spec*.+E16), constituent bien des traductions de la forme *fonctionnaires* (au même titre que la forme *servants* et le segment *civil servants* découverts précédemment), (cf. *Tableau 4*). L'exploration contextuelle s'appuie sur les outils de navigation textométrique de la *carte des sections*. Sur la figure 2, les boutons situés à gauche de la fenêtre de visualisation de la sélection (en forme de mains) permettent de passer, respectivement, à la section suivante/précédente ou l'occurrence suivante/précédente du type étudié. Pour explorer parallèlement les deux volets bilingues du corpus, les sections en correspondance sont liées. Toute sélection dans une fenêtre est répercutée dans l'autre.

volet français	volet anglais
/…/ l'introduction de procédures disciplinaires à l'encontre de **fonctionnaires**, en raison de leur engagement politique /…/, violerait la convention de l'organisation internationale du travail (oit) /…/	/…/ the institution of disciplinary proceedings against *civil* **servants** on account of their political activities /…/ breached international labour organisation (ilo) convention /…/
il s'agissait en fait d'un document destiné aux agents du bvd (*binnenlandse veiligheidsdienst*) et d'autres **fonctionnaires** appelés à accomplir des missions pour lui.	it was in fact a document intended for bvd (*binnenlandse veiligheidsdienst*) staff and other **officials** who carried out work for the bvd**.**
il dénonce les propos tenus lors de la conférence de presse par le ministre de l'intérieur et les hauts **fonctionnaires** de police qui l'accompagnaient.	he complained of the remarks made by the minister of the interior and the *senior police* **officers** accompanying him at the press conference.

Tableau 4. Retours au contexte (extrait)

3.3 *Analyse des résultats*

L'exploration du corpus ***Convention*** à l'aide de la topographie bi-textuelle a permis de découvrir les principales traductions de la forme-pôle *fonctionnaires* (F=49) : *officers* (f_{locale}=10), *officials* (f_{locale}=7) et *servants* (f_{locale}=31). Un léger écart entre la fréquence globale de cette forme-pôle et le cumul des fréquences locales de ces correspondances en anglais montre qu'il existe au moins un contexte pour lequel la traduction n'a pas été identifiée par notre exploration. Nous pouvons affiner nos constats à travers un retour au texte. Il suffit d'écarter toutes les phrases du fragment anglais où la forme *fonctionnaires* est traduite par *officers*, *officials* ou *servants*. Pour ce faire, on procède à la sélection des phrases du fragment anglais dans lesquelles ces trois formes sont absentes ou contenues en nombre inférieur au total d'occurrences de la forme *fonctionnaires* dans la phrase correspondante en français. Cette recherche aboutit à la localisation sur la carte du corpus du couple de phrases suivantes :

volet français	volet anglais
Aux termes de /…/ la loi-cadre sur les **fonctionnaires** des *länder* /…/ seul peut être nommé fonctionnaire celui qui « offre la garantie qu'il prendra constamment fait et cause pour le régime fondamental libéral et démocratique au sens de la loi fondamentale. »	by virtue of /…/ the **civil service** (general principles) act for the *länder*, appointments to the civil service are subject to the requirement that the persons concerned "satisfy the authorities that they will at all times uphold the free democratic constitutional system within the meaning of the basic law".

(Corpus ***Convention***)

Dans ces dernières phrases, la forme *fonctionnaires* a été traduite par *civil service*. Cette équivalence relève de la notion d'*équivalence contextuelle*. Sur le plan sémantique, il s'agit d'unités traductionnelles singulières qui nécessitent un traitement particulier. Il appartient à l'expert humain de s'appuyer sur les blocs alignés pour examiner dans le détail les parallèles et les divergences entre ce type de séquences : *la loi-cadre sur les fonctionnaires ~ the civil service (general principles) act*.

Conclusions

Au terme de cette étude, nous avons défini une approche qui permet d'accéder à la description automatique de relations de correspondance entre des unités polysémiques qui possèdent plusieurs traductions au sein du corpus bi-textuel. Privilégiant le point de vue textométrique, notre approche repose entièrement sur les ressources construites à base de corpus. Appuyée sur l'utilisation de la représentation topographique de textes alignés au niveau de la phrase, cette approche offre à l'utilisateur de nouveaux moyens informatisés pour explorer la structure des équivalences qui se forment au niveau des mots et des syntagmes dans les textes originaux et leurs traductions.

Références

Heiden S. (2002*). Weblex : Manuel Utilisateur. Version 4.1* (intermédiaire). UMR 8503, ENS Lettres et Sciences humaines : http://lexico.ens-lsh.fr/doc/weblex.pdf.

Isabelle P. et Warwick-Armstrong S. (1993). Les corpus bilingues : une nouvelle ressource pour le traducteur. In Bouillon P. et Clas A. (Eds), *La Traductique : Études et Recherches de traduction par ordinateur.* Les Presses de l'Université de Montréal : 288-306.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie.* Slatkine-Champion.

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des JADT 2002* : 403-412.

Lamalle C., Martinez W., Fleury S., Salem A., Fracchiolla B., Kuncova A. et Maisondieu A. (2003). *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation*. Syled-CLA^{2}T, Université de la Sorbonne nouvelle – Paris 3 : http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW.

Lebart L. et Salem A. (1994). *Statistique Textuelle.* Dunod.

Véronis J. (Ed.) (2000). *Parallel Text Processing: Alignment and use of translation corpora.* Kluwer Academic Publishers.

Zimina M. (2000). Alignement de textes bilingues par classification ascendante hiérarchique. In *Actes des JADT 2000* : 171-178.

Zimina M. (2002). Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. In Véronis J. (Ed.), *Lexicometrica*, n° spécial Corpus alignés : http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm.

Lexicométrie et corpus multilingues

Table-ronde

Abstract

Beyond automatic parallel text alignment, which is now well-known of our scientific community, this panel session focuses on how to *extend statistical techniques* in view of *exploring multilingual textual data*. As regards *parallel corpora*, new tools and methodologies have emerged. Processing *comparable corpora* (i.e. made-up of similar texts which are not the translation of one another) is also a significant challenge. Textual statistics for monolingual corpora can be adapted to this new type of data. Furthermore, some corpora are written in *languages* which raise new issues as concerns textual statistics softwares: for example the management of the characters encoding, the corpus tokenisation into sensible word-like units, or the definition of clear and coherent linguistic annotation schemes. International standards have recently been published and others are in preparation. They constitute efficient guidelines for corpus and linguistical ressources encoding. As they deal with the genuine diversity of languages throughout the world, these standards allow the comparability and the reusability of textual data.

Résumé

Par delà l'alignement automatique de corpus, fondamental mais bénéficiant déjà d'une certaine maturité et d'une bonne notoriété, cette table-ronde s'intéresse à l'*extension* des *techniques et applications lexicométriques* en contexte multilingue. Sont bien évidemment concernés les *corpus parallèles*, pour l'analyse et l'exploitation desquels peuvent être proposés de nouveaux outils et méthodes. Sont également en jeux les *corpus comparables* (non alignés), qu'il s'agit d'être capable d'exploiter moyennent l'adaptation de procédures d'analyse statistique jusque là pratiquées sur des corpus monolingues. Sont enfin également considérés les corpus dont la *langue* introduit de nouvelles questions théoriques et techniques pour les logiciels de lexicométrie, par exemple pour le codage des caractères, le découpage en « mots », ou le codage des informations morphosyntaxiques. Des *normes internationales* récentes et en cours d'élaboration guident maintenant le codage des corpus et des ressources linguistiques de tous ordres. Prenant en compte la diversité des langues, elles visent à favoriser la comparabilité et la réutilisabilité des données textuelles.

Mots-clés : corpus parallèles, corpus comparables, corpus en langues orientales, statistique textuelle, spécificités, réseaux de cooccurrences, segments répétés, représentation topographique de textes parallèles, interfaces de navigation, classification automatique, analyse canonique non linéaire, lexiques multilingues, jeux de caractères, découpage en mots, étiquetage linguistique, normalisation.

1. Présentation du thème

Le besoin se fait de plus en plus sentir d'outils appropriés à la consultation et à l'analyse de corpus multilingues, alignés ou non. Complémentairement aux outils d'alignement (utiles pour constituer les corpus) et d'appariement de termes, il s'agit d'outiller des modes de parcours, de synthèse, et d'extractions contextuels, offrant une approche moins purement terminologique et plus textuelle de ces données multilingues.

C'est une perspective de développement particulièrement importante pour les outils actuels de statistique textuelle. En effet, le traitement des corpus multilingues dans un cadre jusqu'à présent monolingue conduit soit à mêler les textes des différentes langues dans les calculs (sauf un calcul de contraste en les définissant comme des parties), soit à les traiter de façon totalement indépendante, aucune de ces deux voies n'étant pleinement satisfaisante. Il existe bien

quelques bons concordanciers pour corpus parallèles, mais ceux-ci n'offrent pas toute la richesse des calculs statistiques mis au point en lexicométrie.

Différents chantiers sont à considérer :

- la prise en compte de jeux de caractères qui sortent de l'ASCII ou de l'isoLatin1 (arabe, chinois,...)
- la prise en compte de modes de segmentation relatifs aux langues : la notion de forme délimitée par deux caractères séparateurs comme approximation du mot n'est pas directement transposable, cf. par exemple le cas des mots composés en allemand, ou des écritures par idéogrammes syllabiques.
- l'adaptation des calculs statistiques et des interfaces aux calculs alignés : traitements sur une langue et affichage des (ou accès aux) correspondances dans l'autre langue ; interrogations croisées ou/et alternées sur les différentes langues ; rôles bien différenciés des différentes langues (par ex., ce qui éclaire le mieux un calcul de concordance sur une langue n'est pas nécessairement un semblable calcul de concordance sur l'autre langue).
- la conception de nouvelles procédures et calculs statistiques, plus particulièrement liés aux besoins d'investigation que peuvent avoir des utilisateurs (enseignants de langues, linguistes, traducteurs, industriels, professionnels de l'information) sur ce type de corpus.

Le but de cette table-ronde est de réunir les personnes de la communauté des statistiques textuelles intéressées par ces questions, afin :

- de débattre sur les perspectives les plus importantes et intéressantes d'extension des techniques lexicométriques à ce terrain des corpus multilingues
- et éventuellement d'organiser les efforts de recherche et de développement, cette rencontre favorisant la prise de contacts pour qu'ensuite se positionnent et se coordonnent les acteurs intéressés.

2. Résumés des interventions

Coordination de la table-ronde : ***Bénédicte Pincemin***

CNRS, LLI – Université Paris 13 – Avenue J.-B. Clément – 93430 Villetaneuse – France
http://www-lli.univ-paris13.fr/ benedicte.pincemin@centraliens.net

2.1. *Mathieu Valette*

INALCO, Centre de Recherche en Ingénierie Multilingue – 2, rue de Lille – 75343 Paris cedex 7 – France
http://crim.inalco.fr/ mvalette@inalco.fr

Encore récente, la linguistique de corpus ouvre de nouveaux champs de prospection tant en matière de recherche (analyse de données attestées) que d'application (ingénierie du document). Mais parce qu'elle repose sur la technologie informatique, une minorité de langues en bénéficient réellement : celles écrites avec l'alphabet latin.

Toutefois, l'encodage d'un grand nombre d'alphabets et la normalisation de la structuration des textes (Unicode, TEI) permet d'envisager le traitement automatique de la plupart des langues écrites et subséquemment, le renouvellement de la problématique comparatiste historiquement définitoire des sciences du langage.

L'Institut National des Langues et Civilisations Orientales, et en son sein le Centre de Recherche en Ingénierie Multilingue, sont particulièrement désireux de participer à la pro-

blématisation de ce champ encore peu exploré. Car l'enjeu n'est pas exclusivement théorique et scientifique, il concerne également l'accès vital des langues orientales au secteur quaternaire (ou « nouvelle économie »).

2.2. *Maria Zimina*

SYLED EA 2290 – Université de la Sorbonne nouvelle Paris 3 – 19, rue des Bernardins – 75005 Paris – France
http://www.cavi.univ-paris3.fr/ilpga/ED/student/stmz/ zimina@msh-paris.fr

L'utilisation des pratiques de *textométrie* pour le traitement automatique de corpus parallèles constitue une piste de recherche prometteuse. Dans le contexte multilingue, les méthodes quantitatives (telles que la *classification automatique*, les *spécificités*, le *calcul des réseaux de cooccurrences*, etc.) apportent un éclairage précieux sur des régularités dans la structuration de discours parallèles. Ces méthodes peuvent être utilisées pour l'appariement automatique des vocabulaires et des fragments de texte. Elles reposent entièrement sur des ressources construites à partir de corpus parallèles eux-mêmes. Les fréquences et les distributions d'unités textuelles (formes, segments répétés, etc.) au sein de chacun des volets multilingues servent de points de repère pour l'identification et l'extraction des correspondances. Appuyée sur l'utilisation de la *représentation topographique* du *bi-texte*, l'approche textométrique permet d'envisager la conception de nouvelles procédures informatiques liées aux besoins d'investigation que peuvent avoir des utilisateurs (linguistes, traducteurs, professionnels de l'information, etc.) sur ce type de corpus.

Martinez W. et Zimina M. (2002). Utilisation de la méthode des cooccurrences pour l'alignement des mots de textes bilingues. In *Actes des JADT 2002*.

Zimina M. (2000). Alignement de textes bilingues par classification ascendante hiérarchique. In *Actes des JADT 2000*.

Zimina M. (2002). Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. In Véronis J. (Ed.), *Lexicometrica*, n° spécial Corpus alignés : http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema6.htm

2.3. *Cyril Goutte*

Xerox Research Center Europe – 6 ch. de Maupertuis – 38240 Meylan – France
http://www.xrce.xerox.com/ Cyril.Goutte@xrce.xerox.com

Pour de nombreuses applications multilingues (par exemple la traduction automatique) il est nécessaire de disposer d'un corpus aligné ou parallèle, typiquement issu de traductions. Dans d'autres domaines, il est difficile de disposer de tels corpus – il faut alors travailler à partir de corpus comparables (Déjean *et al.*, 2002), traitant du même sujet sans pour autant avoir de correpondance un-à-un. Dans ce cadre, on évoquera l'utilisation de techniques d'analyse des données qui sont le prolongement des décompositions spectrales types « Latent Semantic Analysis » utilisées en monolingue. En particulier, on présentera l'utilisation de l'analyse canonique, et son extension au domaine non-linéaire (Lai et Fyfe, 2000), pour l'enrichissement de lexiques bilingues à partir de corpus comparables.

Déjean, Gaussier et Sadat (2002). Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In *Proceedings of COLING'2002*.

Lai et Fyfe (2000). Kernel and nonlinear canonical correlation analysis. *IJNS*, vol. (10/5) : 365-377.

2.4. *Laurent Romary*

LORIA UMR 7503 – Campus Scientifique – BP 239 – 54506 Vandœuvre-lès-Nancy CEDEX – France
http://www.loria.fr/equipes/led Laurent.Romary@loria.fr

Perspectives de normalisation pour les corpus et lexiques multilingues

Afin d'effectuer des recherches de plus en plus fines sur la langue, il est nécessaire de disposer d'une infrastructure stable mais flexible pour représenter différents types de ressources linguistiques : corpus annotés et lexiques. Au delà de la simple représentation de données primaires, pour lesquelles le cadre de la TEI (Text Encoding Initiative) est maintenant bien reconnu, je présenterai les grandes lignes d'action du nouveau comité TC 37/SC 4 de l'ISO qui travaille d'une part sur la modélisation d'annotations multi-niveaux et de ressources linguistiques multilingues. Je présenterai en particulier le projet de mise en œuvre d'un registre de catégories de données en linguistique qui, dans une perspective internationale, devrait permettre de disposer d'un référentiel de comparaison de différents schémas d'annotation ou de bases lexicales.

Je finirai en montrant comment de tels standards devraient nous permettre de mieux échanger des données et des outils au sein des communautés de linguistique et de traitement des langues, en m'appuyant sur l'expérience en cours d'espace libre de gestion de ressources linguistiques pour le français FReeBank.

3. Communications en lien avec cette table-ronde

Plusieurs communications présentées à ces JADT rejoignent la problématique abordée dans cette table-ronde, tout particulièrement :

Bécue M., Pagès J. et Pardo C.-E. Analysis of multilingual free responses.

Deroubaix J.-C. Que faire des corpus multilingues parallèles ? Une expérience.

Deville G., Dumortier L. et Paulussen H. Génération de corpus multilingues dans la mise en oeuvre d'un outil en ligne d'aide à la lecture de textes en langue étrangère.

Corneille et Molière

Table ronde

Dans ses publications et son ouvrage : *Corneille dans l'ombre de Molière*, Dominique Labbé, spécialiste de statistique lexicale, affirme avoir contribué à résoudre une énigme scientifique en déterminant que certains chefs-d'œuvre de Molière (*Le Tartuffe*, *Dom Juan*, *Le Misanthrope*, *L'Avare*) sont en fait l'œuvre de Corneille. L'argument principal de cette découverte est l'étonnante similarité du vocabulaire des plus grandes pièces de Molière et de deux comédies de Corneille. Parce qu'elle a paru ébranler un mythe national, cette affirmation a provoqué un bouleversement qui a dépassé le seul monde des érudits pour atteindre la grande presse et les médias. Les conclusions de D. Labbé sont fondées sur la « distance intertextuelle », instrument statistique développé avec Cyril Labbé. Cette méthode mesure la différence globale de vocabulaire entre deux textes en déterminant leur différence relative en occurrences de mots : plus le nombre est bas, plus il est probable que les œuvres soient du même auteur. Les auteurs précités concluent que la distance lexicale de 16 pièces de Molière avec deux pièces anciennes de Corneille est suffisamment proche de zéro pour prouver que les textes sont de la même main. La polémique suscitée a mobilisé des statisticiens et des spécialistes du théâtre du XVIIe siècle : les uns et les autres estiment que la statistique « ne devrait être utilisée que comme un auxiliaire complétant l'analyse littéraire et les données historiques », comme un « instrument exploratoire avec un objectif de description, mais en aucun cas comme une preuve ». Au cœur du débat se trouve donc la question fondamentale de l'usage qui est fait des instruments statistiques dans le domaine des lettres. Les historiens de la littérature, en outre, apportent dans la question de la paternité des œuvres de Molière des faits et des arguments d'ordre biographique, psychologique, historique, esthétique et stylistique.

Dominique Labbé
CERAT-IEP - Université Pierre Mendès-France – Grenoble
dominique.labbe@iep.upmf-grenoble.fr

Les travaux qui sont à l'origine de ce débat sont techniques, transparents et publiés. Nous n'avons pas une thèse, polémique ou pas, sur le sujet, mais une méthodologie dont un résumé de 20 lignes (demandé un peu tardivement) ne peut rendre compte. Nous pensons plus utile de donner quelque références.

Notre travail a été publié dans une revue internationale à comité de lecture :

> Labbé Cyril, Labbé Dominique. « Inter-Textual Distance and Authorship Attribution Corneille and Molière". *Journal of Quantitative Linguistics*. 8-3. December 2001. 213-231.

La distance intertextuelle a été présentée en français (avec une discussion de ses propriétés) :

> Labbé Cyril, Labbé Dominique. « La distance intertextuelle ». *Corpus*. 2-2003. 95-118.

La méthode a été validée par plusieurs expériences techniquement irréprochables ; elle est utilisée par plusieurs chercheurs. Des publications sont en cours. Deux articles ont déjà paru :

Merriam Thomas. “Intertextual Distances between Shakespeare Plays, with Special Reference to Henry V (verse)”. *Journal of Quantitative Linguistics*. 9-3. December 2002. 260-273.

Merriam Thomas. “An Application of Authorship Attribution by Intertextual Distance in English”. *Corpus*. 2. 2003. 167-182.

Une synthèse concernant l’ensemble du dossier est accessible à :

http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeCorneilleMoliere.pdf

(On trouvera sur cette page une série d’autres documents sur ce thème).

L’Université d’Oxford a mis en ligne toutes les pièces de Corneille, Molière et Racine — normalisées et lemmatisées par nos soins — avec d’autres corpus traités de la même manière, des publications et la plupart de nos logiciels : http://ota.ahds.ac.uk/2466.

Nous n’avons rien à vendre et nous n’utilisons aucune boîte noire.

Jean-Marie Viprey
Archives, Textes, Sciences des Textes (EA 3187)
Université de Franche-Comté

Les statistiques lexicales sont-elles destinées à fournir des preuves ? Ne servent-elles pas plutôt à renforcer ou affaiblir des hypothèses, et surtout à les réorienter ? Une méthode probabiliste ne saurait, à elle seule, prouver qu’un auteur peut être tenu pour responsable de telle ou telle œuvre particulière.

Corneille et Molière sont intervenus dans un même champ littéraire, à des dates proches, et dans un genre très contraignant, la comédie (voire la comédie en vers). Corneille a mis la main à *Psyché*. Molière a écrit *Dom Garcie de Navarre* sous l’influence de la comédie héroïque cornélienne.

À mon sens, les résultats obtenus jusqu’ici accréditent plutôt la thèse communément admise : Molière invente peu à peu, et de façon décisive à partir de 1662 (*L’École des femmes*), une expression théâtrale radicalement nouvelle.

L’analyse lexicale massive, qui ne prend pas en compte les collocations ni *a fortiori* les configurations plus complexes, me paraît insuffisante pour caractériser une écriture. Par exemple, l’examen des interjections employées exclusivement dans les textes signés par Molière manifeste la référence spécifique de ce dernier à la Commedia dell’Arte (à l’exclusion de toutes les pièces de Corneille), ce qui affaiblit notablement la thèse présentée par Dominique Labbé.

Index des auteurs / Authors Index

ÁLVAREZ, Ramón : 42.
AMATO, Silvano : 52.
ANGHELUTA, Roxana : 60.
ARENI, Alessandra : 262.
AREZKI, R. : 816.
BAGAVANDAS, Mappillairaju : 71.
BALBI, Simona : 79.
BARBU, Ana-Maria : 88.
BARET, Philippe V. : 700.
BARTOLETTI, Silvia : 99.
BATTISTI, Nadia : 250.
BEAUDOUIN, Valérie : 107.
BÉCUE, Mónica : 42, 119.
BÉLANGER, Luc : 128.
BERGERON, Jean-Guy : 136.
BERNET, Charles : 148.
BERRIDGE, Damon : 926.
BERRY, Anne : 160.
BESTGEN, Yves : 171, 182, 859.
BEUST, Pierre : 977.
BIBER, Douglas : 15.
BISKRI, Ismaïl : 192.
BOLASCO, Sergio : 202.
BRUGIDOU, Mathieu : 211.
BURNER, Sylviane : 221.
CAJOLET-LAGANIÈRE, Hélène : 764.
CANCEDDA, Nicola : 515.
CAPPADOZZI, Tania : 957.
CAPPELLI, Carmela : 229.
CARBONE, Simona : 241, 804.
CARLI, Renzo : 250.
CASTAGNA, Marco : 1097.
CEMIN, Marisa : 1097.
CHIRUMBOLO, Antonio : 262, 1019.
CLECH, Jérémy : 650.
DAOUST, François : 353, 479.
DEGAND, Liesbeth : 859.
DÉJEAN, Hervé : 515.
DELCHAMBRE, Alain : 682, 1117, 1185.
D'ELIA, Angela : 229.
DELLA FAILLE, Dimitri : 353.
DELLA RATTA – RINALDI, Francesca : 202.
DE MARNEFFE, Marie-Catherine : 270.
DE ROECK, Anne : 282.
DEROUBAIX, Jean-Claude : 295.
DEVILLE, Guy : 304.
DI MEGLIO, Emilio : 52, 79.
DISTER, Anne : 313.
DOAN-NGUYEN, Hai : 325.
DOLCETTI, Francesca : 250.
DOUCET, Antoine : 334.
DRAY, G. : 816.
DROUIN, Patrick : 345.
DUCHASTEL, Jules : 353, 364.
DUFRESNE, Anne : 373.
DUFRESNE, Monique : 479.
DUMORTIER, Laurence : 304.
DUPONT, Pierre : 270.
DURIAU, Vincent J. : 382.
EMIRKANIAN, Louisette : 390.
ERLOS, Frédéric : 399.
EVERT, Stefan : 411.
FAIRON, Cédrick : 182, 423.
FENOGLIO, Irène : 468.
FERNANDEZ TUESTA, Esteban : 824.
FIALA, Pierre : 716.
FOLCH, Helka : 565.
FOREST, Dominic : 434.
FOUQUERÉ, Christophe : 390.
FOUSS, François : 445.
FRANCIS, Brian : 926.
FRANCQ, Pascal : 682, 1117, 1185.
FUJIMURA, Itsuko : 456.
GALLOPEL, Karine : 479.
GANASCIA, Jean-Gabriel : 468.
GARBERO, Alessandra : 99.
GARROUSTE, Damien : 916.
GARTHWAITE, Paul H. : 282.
GAUSSIER, Eric : 515.
GEERAERTS, Dirk : 1079, 1136.
GÉLINAS-CHEBAT, Claire : 479.
GÉRARD, Raphaël : 488.

GILQUIN, Gaëtanelle : 496.
GINEBRA, Josep : 937.
GIULIANO, Luca : 504.
GOUTTE, Cyril : 515, 1203.
GRABAR, Natalia : 1106.
GRASSIA, Maria Gabriella : 524.
GREEVY, Edel P. : 533.
GROSS, Gaston : 545.
GUERRA, Maria : 52.
GUILPIN, Christian : 557.
GUILPIN, Patricia : 557.
HABERT, Benoît : 565.
HEIDEN, Serge : 577.
HEITZ, Thomas : 946.
HENRY, Sandrine : 848.
HO, Ngoc-Diep : 423.
IBEKWE-SANJUAN, Fidelia : 589.
IGELMO, Angel : 601.
ILLOUZ, Gabriel : 565.
ISSAC, Fabrice : 390.
JACOBSON, Michel : 625.
JACQUEMIN, Bernard : 633.
JACQUES, Jean-Marie : 642.
JALAM, Radwan : 650.
JARDINO, Michèle : 661.
JEUNIAUX, Patrick : 60.
JORDÀ, Gabriel M. : 601.
JOYAL, Sylvain : 192.
KABA, Bangaly : 160.
KASTBERG SJÖBLOM, Margareta : 672.
KERVES, Laurent : 182.
KINDT, Bastien : 488.
KODRATOFF, Yves : 946.
KOSSEIM, Leila : 325.
KUMPS, Nicolas : 682.
LABBÉ, Dominique : 136, 1207.
LACOSTE, Francis J. : 364.
LAFOURCADE, Mathieu : 692, 1008.
LALISCIA, Barbara : 1089.
LANTIN, Anne-Catherine : 700.
LAPALME, Guy : 128.
LAURENT, Nathanaël : 642.
LEACOCK, Claudia : 35.
LEBART, Ludovic : 708.
LEBEL, Marie Élaine : 479.
LEBLANC, Jean-Marc : 716.
LEBRAVE, Jean-Louis : 468.
LECOUTRE, Eric : 496.
LEFÈVRE, Alain : 969.
LEJEUNE, Christophe : 726.
LELU, Alain : 737.
LONGOBARDI, Maria : 241.
LONGRÉE, Dominique : 743.
LUONG, Xuan : 743.
MACÉ, Caroline : 700.
MANDRAN, Nadine : 211.
MANGUIN, Jean-Luc : 753.
MANIMANNAM, G. : 71.
MARTEL, Pierre : 764.
MARTINAZIOLI, Sara : 1089.
MASSON, Chantal-Édith : 764.
MATTE-TAILLIEZ, Oriane : 946.
MAUREL, Denis : 776.
MAYAFFRE, Damon : 785.
MAZZIOTTA, Nicolas : 793.
MELLET, Sylvie : 743.
MEUNIER, Jean-Guy : 192, 434.
MISITI, Maura : 804.
MISURACA, Michelangelo : 524.
MITRA, Rudradeb : 60.
MOENS, Marie-Francine : 60.
MOINE, Michel : 211.
MOKRANE, A. : 816.
MONTECOLLE, Silvia : 99.
MUGNAINI, Rogério : 824.
MÜLLER, Jean-Pierre : 831.
NADIF, Mohamed : 160.
NAKAMURA, Takuya : 837.
NAKAO, Hiroshi : 456.
NISCO, Ferdinando : 99.
NUGIER, Sylvaine : 916.
ORRÙ, Luisa : 1089.
OTRANTO TARDELLI, Adalberto : 824.
PAGÈS, Jérôme : 119.
PALLAUD, Berthille : 848.
PARDO, Campo-Elias : 119.
PAULUSSEN, Hans : 304.
PERADOTTO, Anne : 916.
PIÉRARD, Sophie : 859.
PINCEMIN, Bénédicte : 865, 1203.
PINEIRA-TRESMONTANT, Carmen : 874.
PIRON, Sophie : 885.

PIZARRO NOËL, François : 364.
POIBEAU, Thierry : 897.
POMMIER, Jean-Luc : 904.
PONCELET, P. : 816.
PRINCE, Violaine : 692, 1008.
QUATRAIN, Yasmina : 916.
RAKOTOMALALA, Ricco : 650.
RASOLOFO, Yves : 1000.
RAYSON, Paul : 926.
RECCHINI, Emanuela : 99.
REGER, Rhonda K. : 382.
RENDERS, Jean-Michel : 445.
RIBA, Alex : 937.
ROCHE, Mathieu : 946.
ROMANO, Maria Clelia : 957.
ROMARY, Laurent : 1203.
ROUX, Cristelle : 969.
ROY, Thibault : 977.
SAERENS, Marco : 445.
SALEM, André : 986.
SALERNO, Irene : 99.
SALOMON, Annie-Claude : 211.
SANJUAN, Eric : 160, 589.
SANSONETTI, Luiggi : 993.
SARKAR, Avik : 282.
SAVOY, Jacques : 1000.
SCEPI, Germana : 524.
SCHWAB, Didier : 1008.
SELLERI, Patrizia : 1051.
SENSALES, Gilda : 1019.
SIGAYRET, Alain : 160.
SMEATON, Alan F. : 533.
SPEELMAN, Dirk : 1079, 1136.
SZMRECSÁNYI, Benedikt : 1031.
TOMA, Cristina Alice : 1039.
TOMASETTO, Carlo : 1051.
TRÉBUCQ, Stéphane : 1061.
TUERLINCKX, Laurence : 1069.
TUMMERS, Jose : 1079.
TURCHI, Gian Piero : 1089.
TUZZI, Arjuna : 1097.
UCHIDA, Mitsumi : 456.
VALENCIA, Olga : 42.
VALETTE, Mathieu : 1106, 1203.
VANDAELE, Valéry : 1117.
VANGEHUCHTEN, Lieve : 1128.
VAN GIJSEL, Sofie : 1136.
VENANT, Fabienne : 1145.
VERGNE, Jacques : 1157.
VICENS, Carlota : 601.
VIPREY, Jean-Marie : 1165, 1207.
VO TRUNG, Hung : 1177.
WALLEMACQ, Anne : 642.
WARTEL, David : 1185.
YVON, François : 107.
ZIMINA, Maria : 1195, 1203.

Index des mots-clés / Keywords Index

Acquisition automatique de la terminologie : 345.
Acquisition du langage : 993.
Acteurs, Actors : 1061.
Adjectif épithète : 456, 753.
Adjectifs : 241.
Adjectival inflection : 1079.
Adjectives : 241.
Aggettivi : 241.
Alceste : 250, 904, 1051.
Algorithme centres de gravité mobiles : 661.
Alignement de corpus multilingues : 304.
Alignement lexical : 1195.
Amorces de mots : 848.
Analisi delle co-occorrenze : 250.
Analisi delle corrispondenze semplici : 99.
Analisi emozionale del testo (AET) : 250.
Analisi lessicale : 804.
Analisi statistica di dati testuali : 1051.
Analisi testuale : 99.
Analyse : 764, 1069.
Analyse arborée : 785.
Analyse automatique d'un discours spécialisé : 837.
Analyse canonique non linéaire : 1203.
Analyse de contenu : 364, 1061.
Analyse de corpus : 885, 993.
Analyse de données textuelles : 708.
Analyse de fréquence : 345.
Analyse de texte par ordinateur : 353, 364.
Analyse de textes : 479.
Analyse du discours : 211, 642, 1039.
Analyse électorale : 211.
Analyse en composantes locales : 737.
Analyse en composantes principales : 977.
Analyse factorielle des correspondances : 364, 904.
Analyse lexicale : 211, 479.
Analyse linguistique de surface : 897.
Analyse multi-dimensionnelle : 1165.
Analyse sémantique : 969, 1061.
Analyse sémantique latente : 171, 859.
Analyse statistique textuelle de corpus : 373.
Analyse syntaxique : 1061.
Analyse thématique : 211, 434.
Analyses arborées : 743.
Anaphore : 859.
ANOVA : 71.
Apprentissage : 192.
Apprentissage lexical : 692.
Approche inductive et itérative : 479.
Approche semi-automatique : 192.
Approches statistiques : 897.
Arabe : 1069.
Arbre non équilibré : 661.
Article : 456.
Association de mots : 682.
Association rules : 79.
Atteggiamenti degli adolescenti : 804.
Author style statistics : 71.
Authority : 1117.
Automated scoring : 35.
Bag-of-words : 79.
Banque Centrale Européenne : 373.
Base de données textuelles : 764.
Base lexicale sémantique : 1008.
Bibliometrics : 824.
Biological entity recognition : 515.
Bootstrap : 42, 708.
Brazilian : 824.
Brigate Rosse : 262.
Calcul du sens : 1145.
Candide : 726.
CAQDAS : 726.
Carte de lien : 726.
Cartographie : 1061.
Cartographie d'information : 816.
Cartographie thématique de corpus : 977.
Catégorisation : 202, 434, 479.
Catégorisation hiérarchique ascendante : 977.
Categorizzazione sociale : 262.

Causative constructions : 496.
Centralité : 737.
Champ sémantique : 642.
Champs stylistiques : 601.
Change point estimation : 937.
Characteristic textual units : 1097.
Chartes médiévales : 793.
Chirac/Jospin : 785.
Chi-squared test : 926.
Citation : 824.
Classes d'objets : 545.
Classification : 434, 1177.
Classification automatique : 737, 1106, 1203.
Classification non supervisée : 661.
Classification of manuscripts : 700.
CLEF collection : 650.
Cluster : 726.
Cluster analysis : 250.
Clustering : 60, 160, 589.
Coding list adaptations : 957.
Communauté française de Belgique : 313.
Comparazione tra sub-testi : 241.
Competitive intelligence : 589.
Computer-aided text analysis : 382.
Concepts : 1061.
Conceptual vectors : 692, 1008.
Concordance : 577.
Concordances : 764.
Connaissances indépendantes de tout domaine : 171.
Construction hypothesis : 1008.
Content analysis : 1061.
Contesto discorsivo : 1051.
Contextes : 764.
Contextual co-occurrence : 816.
Contraintes syntaxiques : 885.
Cooccurrence : 577, 682, 726, 1165.
Co-occurrence contextuelle : 816.
Cooccurrence statistics : 411.
Cooccurrences : 716.
Coreference resolution : 60.
Corneille : 148.
Corporate governance : 1061.
Corpus : 345, 764.
Corpus comparables : 1203.
Corpus d'entrevues de groupes : 479.
Corpus économique : 837.
Corpus en langues orientales : 1203.
Corpus français écrit : 753.
Corpus linguistics : 865, 926, 1128.
Corpus multilingues : 295.
Corpus oraux : 625.
Corpus parallèles : 1195, 1203.
Corpus profiling : 282.
Corpus representatif : 136.
Corpus sur le Web : 353.
Correspondances de traduction : 1195.
Correspondence analysis : 42, 504, 1019.
Corrispondenze lessicali : 241, 804.
Courbes de distribution : 743.
C-rater : 35.
Critère de Kolmogorov : 557.
Cross language open-ended questions : 119.
Cultura locale : 250.
Daily activities : 957.
Decision explorer : 1061.
Decision tree : 496.
Découpage en mots : 1203.
Découverte des mots vides : 1157.
Degré de collocation : 456.
Densité : 737.
Dépression : 221.
Derivation : 633.
Désambiguïsation : 1145.
Détection : 1106.
Diachronie : 1165.
Dictionnaire dérivationnel : 488.
Dictionnaire électronique : 488.
Discours politique : 295, 716, 785, 874.
Discours scientifique : 1039.
Discours scolaire : 904.
Discrimination techniques : 270.
Discriminative : 515.
Discursif : 456.
Discursive processes : 1089.
Disfluences, Disfluencies : 848.
Distance de Hellinger : 737.
Distance intertextuelle : 743, 785.
Diversity indexes : 937.
Divulgative publications : 1089.

Dix-septième siècle : 148.
Domanda e offerta : 250.
Doxa néo-libérale : 373.
DTD : 764.
Dutch : 1079, 1136.
Dynamique : 904.
Eigenvalues : 42.
Entités nommées : 182.
Entretiens sociologiques : 136.
Entropie : 661.
E-rater : 35.
Espace conceptuel : 682.
Espace sémantique : 1145.
Etimologia : 250.
Étiquetage : 488, 716, 776.
Étiquetage linguistique : 1203.
Étiquetage morphosyntaxique : 865.
Étiquetages : 716.
EU ideology : 373.
European Central Bank : 373.
Évaluation : 202, 916.
Expectation-maximization algorithm : 270.
Extraction de question : 128.
Extraction de synonymes : 423.
Extraction de syntagmes : 1106.
Extraction d'information : 182.
Fatware : 793.
Féminisation : 313.
Figement : 545.
Filtrage : 1106.
Filtre sémantique : 423.
Financial markets : 373.
Fisher kernels : 515.
Flot maximal : 423.
Flussi migratori : 99.
Fonctions lexicales : 1008.
Formazione professionale : 250.
Forme : 1069.
Forme caratteristiche : 241.
Forme specifiche : 804.
Formes 1re et 2e personnes : 221.
Fouille de textes : 816.
Français : 456.
Français oral et spontané : 848.
France : 136.
Function words discovery : 1157.
Gender : 1019.
Generative : 515.
Genèse textuelle : 468.
Genres littéraires : 672.
Gestion des relations avec la clientèle : 128.
Gouvernance : 1061.
Gouvernement d'entreprise : 1061.
Grammaire locales : 837.
Grammar induction : 1157.
Graph decomposition : 160.
Graphe : 1117.
Graphes : 737.
Grec ancien : 488.
Héritage : 545.
Hierarchical classification : 957.
Homogénéité et spécificité : 601.
Homogeneity : 282.
Homologies : 468.
Homonymie : 565.
Hub : 1117.
Hyperbase : 865, 1106.
Hyperliens : 1117.
Hyperonymie, Hyperonymy : 692.
Hypertexte : 577.
Hypothèses de constructions : 1008.
Ideologia : 262.
Idéologie communautaire (i.e. de l'Union européenne) : 373.
Immagine : 250.
Inconscio : 250.
Individuation socio-linguistique : 785.
Information : 1185.
Information cartography : 816.
Information géographique : 969.
Information retrieval : 325, 334, 390, 399, 824, 1000.
Information structuring : 633.
Informatisation : 764.
Interaction verbale adulte/enfan : 993.
Interazione : 1051.
Interface : 865.
Interface HTML : 353.
Interfaces de navigation : 1203.
Interviste semi-strutturate : 1051.

Intranet : 399.
Ispell : 831.
Jeux de caractères : 1203.
K-means : 270.
Labeling : 488, 716, 776.
Langage de balisage de texte : 625.
Language identification : 650.
Language variation : 1136.
Langue de spécialité : 345.
Latin : 743.
Le Clézio : 672.
Lecture et analyse de textes assistées par ordinateur : 434.
Lemmatisation : 136, 171, 672, 865, 1069.
Lemme : 488, 1069.
Lexical : 456.
Lexical analysis : 1019.
Lexical collocation : 1079.
Lexical functions : 1008.
Lexical learning conceptual vectors : 692.
Lexical semantic disambiguation : 633.
Lexical statistics : 411.
Lexical tables : 42.
Léxico-estadística : 1128.
Lexicogramme : 577.
Lexicogramme récursif : 577.
Lexicogrammes : 716.
Lexicographie : 399.
Lexicología, Lexicologie, Lexicology : 1128.
Lexicométrie : 399, 672, 708, 874, 993.
Lexique : 488.
Lexique-grammaire : 837.
Lexiques multilingues : 1203.
Lexter : 1061, 1106.
Liens : 1117.
Linguaggio : 262.
Linguistic relations : 160.
Lingüística aplicada a la adquisición de una lengua extranjera : 1128.
Lingüística de corpus : 865, 926, 1128.
Linguistics : 88.
Linguistique de corpus : 313, 865, 926, 1128.
Linguistique de terrain : 625.
Linguistique quantitative : 148.
Literature review : 1019.
LNRE models : 411.
Logiciel : 642.
Logiciel anthropocentré : 977.
Logiciel en ligne d'apprentissage des langues étrangères : 304.
Logiciels : 916.
Logistic regression : 1079.
Log-likelihood ratio : 1079.
Log-likelihood test : 926.
Logométrie : 785.
Loi binomiale : 557.
Loi de Poisson : 557.
LSI : 52.
Machine learning : 533, 650.
Macroeconomic pretexts : 373.
Manie : 221.
Mapping : 1061.
Marchés financiers : 373.
Marqueur de la structure : 859.
Mental illness : 1089.
Mesures de voisinage : 743.
Minimal separators : 160.
Mixture models : 515.
Modèles de langage : 107.
Modelli culturali : 250.
Modération salariale : 373.
Molière : 148.
Monetary stability : 373.
Morphème : 488.
Morphologie : 1069.
Morpho-syntactic analysis : 589.
Mots experts : 1185.
Mots inconnus : 776.
MUB model : 229.
Multi-agents systems : 1008.
Multidimensional scaling : 382.
Multilingual NLP : 1157.
Multilingual text categorization : 650.
Multilinguïsation : 1177.
Multilinguisme : 192, 1177.
Naïve Bayes : 270.
Natural language learning : 1157.
Natural language processing : 692.
Neoliberal doxa : 373.
Newsgroups : 504.

N-grams : 192.
N-grams representation : 650.
NLP : 776.
Nom de métier : 313.
Nom générique : 313.
Nom spécifique : 313.
Noms propres : 776.
Non symmetrical correspondence analysis : 524.
Normalisation : 1203.
Ontologie, ontology : 969, 1061.
Opérations linguistiques : 468.
Oral and spontaneous French : 848.
Organisation documentaire : 1185.
Organisation informationnelle : 1039.
Organisation topicale : 1039.
Organisations : 642.
Paradoxes : 642.
Parole dense : 250.
Parsing : 633.
Partenaires sociaux : 373.
Partially labelled data : 515.
Percolation : 737.
Pertinence thématique et informationnelle : 390.
Petit : 456.
Philologie : 1069.
Phylogeny : 700.
PLS Regression : 52.
Poids : 456.
Points de vue : 1185.
Political discourse : 295, 716, 785, 874.
Politics : 504.
Politique linguistique : 313.
Polysémie : 545, 565, 1145.
Populations parentes : 557.
Positive and negative : 202.
Pré-analyse de corpus : 897.
Prediction : 496.
Préférences sélectionnelles : 885.
Prétextes macro-économiques : 373.
Procustes analysis : 42.
Proper names : 776.
Prospéro : 726.
Protocole : 916.
Psychose bipolaire : 221.
Québec : 136.
Question ouverte : 211.
Question-answering : 128, 325, 633.
Question-réponse : 128, 325, 633.
R : 831.
Racine : 148, 1069.
Random text : 411.
Rankings : 229.
Rapidité d'exécution : 1185.
Rapport à l'autre : 221.
Rappresentazioni : 241.
Rappresentazioni di genere : 241.
Recherche aléatoire : 661.
Recherche de ressources clés : 1000.
Recherche de service en-ligne : 1000.
Recherche d'information : 325, 334, 390, 399, 824, 1000.
Récursivité : 488.
Redondance : 764.
Référence numérique : 128.
Référent : 313.
Référentiel terminologique : 399.
Regression analysis : 1136.
Repeats : 848.
Repérage de points de vue : 565.
Répétitions : 848.
Réponse automatisée au courriel : 128.
Representation : 1061, 1061.
Représentation conceptuelle : 1039.
Représentation du sens : 1008.
Représentation topographique de textes parallèles : 1203.
Représentations textuelles : 793.
Representative term : 816.
Réseaux de cooccurrences : 1203.
Réseaux sémantiques : 399, 692.
Ricerca-intervento : 250.
Rimes : 148.
Saint-Exupéry : 601.
SATO-XML : 353.
SciELO Project : 824.
Second language acquisition research : 1128.
Sectorial publications : 1089.
Segmentation : 79, 1177.

Segmentation automatique : 171, 859.
Segments répétés : 1203.
Semantic analysis : 1061.
Semantic lexical base : 1008.
Semantic networks : 692.
Sémantique distributionnelle : 565.
Sémantique interprétative : 1106.
Sémantique textuelle : 977.
Sense representation : 1008.
Séquences : 468.
Séries textuelles chronologiques : 874.
Similarité de mots : 423.
Similarité, similarity : 682, 824.
Simpson index : 937.
Social partners : 373.
Social psychology : 1019.
Social representations : 1019.
Sous-catégorisation : 885.
Spécificités : 345, 1203.
Spoken advertising language : 1136.
Stabilité monétaire : 373.
Statistical analysis of the textual corpus : 373.
Statistical phrases : 334.
Statistics : 88.
Statistics for terminological purposes : 1128.
Statistique lexicale : 136, 295.
Statistique textuelle : 601, 865, 986, 993, 1106, 1203.
Stemmatology : 700.
Structuration : 764.
Structure latente de textes : 661.
Stylistique : 148.
Stylistique de corpus : 672.
Stylométrie, Stylométry : 107, 672, 937, 1031.
Support vector machines : 515, 533.
Switch regression model : 937.
Synonyms : 229.
Synonymy : 633.
Syntactical phrases : 334.
Syntaxe : 390.
Syntaxic analysis : 1061.
Systèmes multi-agents : 1008.
Tagging : 776.
TAL : 776.
Taltac : 1097.
Temps verbaux : 743.
Tendenze demografiche : 804.
Term distribution : 282.
Term frequency : 524.
Terme : 1039.
Terme représentatif : 816.
Termes complexes : 192.
Terminographie : 399.
Terminología, Terminologie, Terminology : 160, 345, 946, 1128.
Terminological variation : 589.
Testimoni privilegiati : 804.
Text categorisation : 79, 533.
Text corpora : 1031.
Text encoding : 1031.
Text expansion : 633.
Text mining : 52, 160, 202, 334, 589, 816, 916.
Text retrieval : 52.
Texte hétérogène : 1177.
Texte homogène : 1177.
Texte idéologique : 1106.
Texte multilingue : 1177.
Textes parallèles : 986.
Textométrie : 986, 1195.
Texts visualisation : 119.
Textual analysis : 504.
Textual classification : 1031.
Textual data : 957.
Textual data analysis : 524, 1089.
Textual similarity : 119.
Textual statistics : 824.
Textual-approach : 1097.
Théâtre français : 148.
Théorie : 1061.
Théorie de l'information : 661.
Theory : 1061.
Time use : 957.
Topographie bi-textuelle : 1195.
Topologie textuelle : 743.
Traitement automatique : 545.
Traitement automatique du langage naturel : 692.
Traitement de la langue naturelle : 1177.
Traitements multilingues : 1157.
Transductive inference : 515.

Translation effects on text categorization : 650.
Translation equivalence : 88.
Trend mapping : 589.
Tropes : 969, 1061.
Ttda : 831.
Two-sample t-statistic and factor analysis : 71.
Typologie textuelle : 672.
Ultrametrics : 160.
Unknown words : 776.
Usage : 969.
Valence affective : 182.
Validation : 708.
Variables tensives : 904.
Variation de position : 753.
Variational linguistics : 1079.
Variations linguistiques : synchronie et diachronie : 557.
Vecteurs conceptuels : 692, 1008.
Verbe *entendre* : 885.
Verifica : 250.
Vers : 148.
Vocabulaire : 1165.
Vocabulary comparison : 1097.
Voisins réciproques : 737.
Wages moderation : 373.
War : 504.
Web : 390, 1000.
Weblex : 577, 865.
Weight : 456.
Word alignment : 88.
Word fragments : 848.
Word frequency : 926.
Word sense disambiguation : 270.
Word senses' identification : 229.
Work : 1019.
XML : 764, 793.
Zipf-Mandelbrot law : 411.