TEXTE ET CORPUS :

Actes des quatrièmes Journées de la Linguistique de Corpus

Sous la direction de Geoffrey Williams

SOMMAIRE

TEXTE ET CORPUS : 1

SOMMAIRE 3

PRÉSENTATION DES AUTEURS 5

Introduction
Geoffrey Williams 7

CORPUS : DE LA CONSTITUTION AUX APPLICATIONS 11

Stratégie de constitution de corpus oraux transcrits (II) : pistes méthodologiques adoptées pour la crétion d'un corpus thématique dans le cadre du projet OuRAL
Nathalie Gasiglia 13

Transcription prosodique pormalisée au sein du projet PFC (Phonologie du Français Contemporain) : l'état d'un chantier
Anne Lacheret-Dujour, Chantal Lyche 24

Le Web entre corpus linguistique et aberration linguistique : le cas des pronoms anglais de première personne
Antoine Consigny 32

Utilisation du Web comme méga-base de données linguistique : application à la traduction de cooccuurences lexicales françaises-anglaises
Chrystel million & Stéphanie Léon 43

Contribution de la lexicométrie à l'approche sémantique des corpus. La forme « texte » dans un corpus des « études littéraires »
Driss Ablali 51

La sémantique lexicale et les genres : analyse systématique d'un corpus québécois
Margareta Kastberg Sjöblom 63

Organisation informationnelle du discours oral narratif en F.L.E. : étude de quelques cas en milieu homoglotte
Hanitra Maury 73

Constitution et exploitation pédagogique de ressources linguistiques pour un enseignement / apprentissage du discours rapporté
Fanny Rinck, Thomas Lebarbé 83

THÉMATIQUE LINGUISTIQUE 95

Quelques adverbes modalisateurs dans le français parlé L2 et L1 : insertion dans le *paragraphe oral*
Victorine Hancock 97

Annotation de la coréférence entre expressions référentielles
Olivier Tardif 105

Deux indices pour l'étude des marqueurs de la continuité thématique dans de grands corpus
Sophie Piérard & Yves Bestgen **111**

Principe de composition sémantico-discursif des collocations lexicales : de l'arbitraire au motivé ?
Estelle Dubreil **121**

Constructions de classes sémantiques non présupposées
Nicolas Gagean & François Rousselot **133**

Cartographie de corpus pour l'étude de métaphores conceptuelles
Thibault Roy, Stéphane Ferrari, Pierre Beust **143**

Les corpus spécialisés au service de la sémantique quantitative : la polysémie du français technique
Ann Bertels **153**

Analyse distributionnelle de corpus de langue générale et synonymie
Edith Galy & Didier Bourigault **163**

Étude simultanée de la synonymie et de l'antonymie en contexte par analyse distributionnelle contrainte
Jean-Luc Manguin, Itsuko Fujimura, Guillaume Jacquet, Fabienne Venant **175**

ORIENTATION MACHINE **185**

Une plate-forme logicielle et une démarche pour la validation de ressources linguistiques sur corpus : application à l'évaluation de la détection automatique de cadres temporels
Stéphane Ferrari, Frédérik Bilhaut, Antoine Widlöcher, Marion Laignelet **187**

Alignement d'unités textuelles de taille variable
Emmanuelle Giguet & Marianna Apidianaki **197**

Une approche de l'expression orale des émotions : étude d'un corpus réel
Solange Hollard, Mutsuko Tomokiyo, Denis Tuffelli **207**

Extraction automatique de termes à partir de textes polonais
Anna Kupść **213**

Corpus de dialogue homme/machine : annotation sémantique et compréhension
Jeanne Villaneau & Sylvain Lamprier **221**

Bases de données et traductions : aspects quantitatifs et qualitatifs
Yvon Keromnes **231**

Appréciations sur la fiabilité d'un étiquetage morphosyntaxique généré automatiquement et présentation du travail lié à son enrichissement
Vassil Mostrov **239**

PRÉSENTATION DES AUTEURS

Driss ABLALI	Université de Franche-Comté
Marianna APIDIANAKI	LATTICE UMR 8094
Ann BERTELS	Université catholique de Louvain, ILT - QLVL
Yves BESTGEN	Université catholique de Louvain, Fonds national de la recherche scientifique
Pierre BEUST	Université de Caen, GREYC – CNRS (UMR 6072)
Frédérik BILHAUT	Université de Caen, GREYC – CNRS (UMR 6072)
Didier BOURIGAULT	Université Toulouse le Mirail, Equipe de recherche en syntaxe et sémantique (CNRS)
Antoine CONSIGNY	Université Marc Bloch, Strasbourg 2. EA 1339.
Estelle DUBREIL	Laboratoire informatique de Nantes Atlantique – LINA FRE CNRS 2729
Stéphane FERRARI	Université de Caen, GREYC – CNRS (UMR 6072)
Itsuko FUJIMURA	Université de Nagoya
Nicolas GAGEAN	LGECO – INSA
Edith GALY	Université Toulouse le Mirail, Equipe de recherche en syntaxe et sémantique (CNRS)
Nathalie GASIGLIA	Université Lille 3. UMR Savoirs, Textes, Langage du CNRS (8163)
Dirk GEERAERTS	Université catholique de Louvain, QLVL
Emmanuel GIGUET	Université de Caen, GREYC – CNRS (UMR 6072)
Victorine HANCOCK	Université de Stockholm
Solange HOLLARD	Laboratoire CLIPS, équipe GEOD
Guillaume JACQUET	Université de Paris 6, LATTICE – CNRS
Margareta KASTBERG SJÖBLOM	ILF-CNRS, Bases Corpus et Langage (UMR 6039)
Yvon KEROMNES	Université de Metz, ATILF – CNRS Nancy.
Anna KUPŚĆ	CNRS, LORIA (UMR 7503) et Académie Polonaise des Sciences, Institut d'informatique.

Thomas LEBARBÉ	Université Stendhal - Grenoble 3, LIDILEM (EA 609)
Anne LACHERET-DUJOUR	Institut Universitaire de France. Laboratoire CRISCO
Marion LAIGNELET	Université de Toulouse le Mirail, ERSS – CNRS (UMR 5610)
Sylvain LAMPRIER	Université de Bretagne Sud, M2R IHM
Stéphanie LÉON	Université de Provence, équipe DELIC.
Chantal LYCHE	Universitetet i Oslo/tromsø
Jean-Luc MANGUIN	Université de Caen, CRISCO - CNRS
Hanitra MAURY	Université de Bretagne Sud, CRINI (EA 1192) Nantes
Chrystel MILLON	Université de Provence, équipe DELIC.
Vassil MOSTROV	UMR Savoirs, Textes, Langage (UMR 8163 du CNRS) - Université Lille 3
Sophie PIÉRARD	Université catholique de Louvain, Fonds national de la recherche scientifique
Fanny RINCK	Université Stendhal - Grenoble 3, LIDILEM (EA 609)
François ROUSSELOT	LGECO – INSA
Thibault ROY	Université de Caen, GREYC – CNRS (UMR 6072)
Dirk SPEELMAN	Université catholique de Louvain, QLVL
Olivier TARDIF	France Télécom / Université de Provence, DELIC
Fabienne VENANT	Université de Paris 6, LATTICE – CNRS
Mutsuko TOMOKIYO	Laboratoire CLIPS, équipe GETA
Denis TUFFELLI	Laboratoire CLIPS, équipe GEOD
Jeanne VILLANEAU	Université de Bretagne Sud, Valoria
Antoine WIDLÖCHER	Université de Caen, GREYC – CNRS (UMR 6072)

INTRODUCTION

Geoffrey Williams
Université de Bretagne Sud

La linguistique de corpus, la linguistique sur corpus, la lexicométrie, l'analyse statistique de données textuelles, le traitement automatique de la langue, et probablement d'autres... Leur point commun est l'emploi d'un corpus, ou d'une ressource numérique.

Sinclair a fourni une définition stricte du terme corpus comme :

> « un assemblage de textes stockés dans un format électronique et qui ont été sélectionnés en fonction de critères externes afin de représenter, dans la mesure de possible, une langue ou une variété de langage comme une source de données pour la recherche linguistique » (Sinclair 2005 :16)[1].

Une telle définition est essentielle pour pouvoir distinguer le corpus employé en linguistique de corpus de la multitude d'autres définitions possibles de ce mot. Elle est le fruit de presque 50 ans de recherche sur des corpus puisque, pour bien comprendre un corpus, ses avantages et ses limites, il faut soi-même l'élaborer et le pratiquer. Un corpus est une question de représentabilité, et surtout de justification. Il faut comprendre les contextes de culture et de situation qui ont engendrés les textes. Nul besoin d'expliciter l'importance de l'environnement textuel pour l'école contextualiste, dont est issu le principal courant en études de corpus au niveau international, pour elle cela va de soi. Les ensembles textuels qui ne remplissent pas les critères de Sinclair doivent être considérés comme des ressources numériques, mais là aussi il faut pouvoir justifier des critères de sélection.

Une telle définition est pourtant également trop restrictive pour notre besoin ici. Le terme ressource numérique est vaste et peut inclure de nombreuses ressources qui n'ont rien à voir avec un ensemble textuel. Nous préférons donc considérer le mot corpus comme hypéronyme et hyponyme. Au niveau supérieur nous pouvons facilement nous mettre d'accord sur l'idée qu'un corpus est un vaste ensemble de données textuelles sous format numérique. Deux notions clés donc, des **données textuelles**, et donc pas des phrases isolées de leur contexte d'emploi ou inventées ou sélectionnées par le linguiste, et **dans un format électronique**. L'autre point commun des recherches ici présentées est l'emploi des outils informatiques dans l'analyse de nos textes. Il ne s'agit pas d'employer des outils bureautiques qui ne sont pas faits pour analyser des textes, mais d'utiliser des outils spécifiques dédiés.

L'emploi des outils informatiques nous amène à un autre facteur essentiel dans l'étude de corpus, la taille des données. Avec l'accès facile à l'Internet, les corpus deviennent de plus en plus grands. Cette taille importante peut traduire la simple compilation d'un tout venant dont il est impossible de tirer des conclusions quand il est cependant nécessaire d'apprivoiser les données. L'inverse n'est également pas bon : si le corpus est trop petit, ce

---

[1] "*A corpus is a collection of pieces of language text in electronic format, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research*".

n'est pas la peine d'utiliser un ordinateur. À part des extrêmes, la taille du corpus est en fonction des besoins de l'étude. Il va de soi qu'un corpus d'écrit sera nettement plus important qu'un corpus d'oral, les contraintes de transcription sont loin d'être négligeables. Pour les écrits, il est ridicule d'utiliser le mot corpus pour seulement un ou deux textes, simplement parce que le mot est à la mode. L'analyse de corpus est largement qualitative, l'étude de corpus fait appel à des données quantitatives. Les deux approches ne sont pas en concurrence, elles sont complémentaires. Aucun de travail sur corpus ne peut être totalement quantitatif, il doit toujours y avoir un va-et-vient entre le quantitatif et le qualitatif, au risque sinon de se limiter à ne faire que compter des choses sans prendre compte du fait que la langue est vivante.

Dans le recueil de textes qui suit, nous avons donc deux points communs : l'emploi d'un corpus dans le sens générique et l'emploi des outils informatiques, afin d'analyser des aspects de la langue naturelle. Pour le reste nos objectifs et applications peuvent être extrêmement variés. C'est la variété qui fait la richesse de la discipline et des Journées de la Linguistique de Corpus.

Les JLC sont un lieu de rencontre entre praticiens de l'analyse informatisée de la langue. Dorénavant bisannuelles, ces journées servent de point de ralliement entre chercheurs confirmés, jeunes chercheurs et doctorants. Il n'y a pas d'a priori, pas de thèmes imposés. Le but est de confronter ses idées et ses analyses aux idées des autres. C'est comme cela que l'on avance. Cependant, quand il s'agit de publier les actes un problème s'impose, comment ordonner au mieux les contributions ?

Une solution est d'employer une classification alphabétique par nom d'auteur. Ce n'est peut-être pas la meilleure, mais c'est celle adoptée par des dictionnaires et beaucoup d'autres outils de référence. C'est une solution qui a aussi l'avantage d'être objective. Toute autre classification est forcement subjective. On peut trier par type de corpus, écrit ou oral ; par approche, davantage TAL ou corpus ; par thème, terminologie, phonologie… Dans ce dernier cas la liste sera longue. Finalement nous avions décidé de regrouper des textes par grand thème ou par application. Les grands thèmes sont, par exemple, les problématiques de constitution de corpus ou les analyses sémantiques, et les applications comme l'enseignement du Français Langue Etrangère. Les articles de ces grands groupes sont présentés par ordre alphabétique.

Notre principal point commun est le corpus, il est donc normal que plusieurs textes traitent de la problématique de constitution de corpus et de leur emploi dans des situations applicatives précises. Ainsi nous avons six textes. Nathalie Gasiglia est concerné par la constitution d'un grand corpus d'oral. La question est importante puisque malgré le travail sur l'oral, les corpus restent petits et dispersés, la constitution d'un grand ensemble transcrit sera un grand pas en avant pour la recherche. Toujours dans l'oralité, nous trouvons les corpus de Phonologie du Français Contemporain (PFC). Anne Lacheret-Dujour et Chantal Lyche nous décrivent l'état de ce chantier de grande envergure. Pour l'écrit, le web représente une ressource beaucoup plus facile d'accès, mais semée de dangers. Antoine Consigny explore des possibilités d'une analyse des pronoms utilisant le web comme ressource, et la grammaire systémique et fonctionnelle comme modèle. Les travaux de Chrystel Millon et Stéphanie Léon utilisent aussi le web, mais cette fois-ci pour une analyse collocationnelle dans un objectif de traduction automatique. Driss Ablali et Margareta Kastberg Sjöblom ont une approche plus lexicométrique avec l'outil Hyperbase.

Dans le cas d'Ablali, il s'agit d'un travail sur les genres dans un corpus de français académique ; le travail de Kastberg Sjöblom traite d'un corpus de québécois.

Le Français Langue Etrangère est représenté par deux articles. L'étude de Hanitra Maury concerne l'analyse de locuteurs non-francophones dans le cadre des critères d'évaluation, le travail de Fanny Rinck et de Thomas Le Barbé est orienté vers la constitution de ressources linguistiques pour une application de TAL dans l'enseignement du français.

La thématique linguistique est beaucoup plus diverse. La communication de Victorine Hancock est la seule qui traite de l'oral dans une étude des adverbes modalisateurs dans le français parlé. Les huit autres textes utilisent des corpus écrits. Olivier Tardif étudie le problème de référence anaphorique dans une ressource numérique, *Le Monde*. Dans le texte de Sophie Piérard et Yves Bestgen, nous avons une étude des marqueurs de la continuité thématique, cette fois-ci dans un corpus littéraire. Deux textes traitent de la collocation, question centrale en linguistique de corpus. L'article d'Estelle Dubreuil est une analyse sémantique des constituants dans un corpus de textes scientifiques ; les travaux de Nicolas Gagean et de François Rousselot sont applicatifs et traitent de la construction automatique de classes. Quatre textes traitent plus proprement de questions de sémantique lexicale. Thibault Roy *et al.*, travaillent sur des métaphores conceptuelles ; Anne Bertels *et al.*, sur la polysémie avec un corpus de français technique. Deux textes ont un regard sur la synonymie dans des corpus de langue générale. L'approche d'Edith Galy et Didier Bourigault est davantage TAL, avec une étude distributionnelle, quand celle de Jean-Luc Manguin *et al.*, qui est plus linguistique dans son orientation. Les deux font référence au dictionnaire des synonymes du CRISCO à Caen.

Cinq textes ont une orientation plus machine. Celui de Stéphane Ferrari *et al.*, concerne une plate-forme pour la validation de ressources linguistiques sur corpus, celui de Jeanne Villaneau et Sylvain Lamprier est un outil pour aider à l'annotation sémantique dans le cadre de l'interaction homme-machine. Solange Hollard *et al.*, présente un travail sur l'expression des émotions dans un corpus oral. Les deux derniers textes de cette section nous amènent vers la traduction. Emmanuelle Giguet et Marianna Apidianaki travaillent sur le problème d'alignement de textes dans des corpus bilingues et multilingues ; Anna Kupść présente le problème d'extraction terminologique sur un corpus en langue polonaise.

Une richesse des approches, une richesse des contenus. La vocation des Journées de la Linguistique de Corpus est précisément de présenter cette richesse et de provoquer des échanges. La publication en-ligne permet de partager ces idées librement, et de continuer à stimuler le débat sur la nature des corpus et les très nombreuses applications.

RÉFÉRENCES

Sinclair John, 2005. 'corpus and text – Basic principles'. in Wynne Martin (ed.), *Developing linguistic corpora : a guide to good practices*. Oxford : Oxbow Books.

CORPUS :

DE LA CONSTITUTION AUX APPLICATIONS

STRATÉGIE DE CONSTITUTION DE CORPUS ORAUX TRANSCRITS (II) : PISTES MÉTHODOLOGIQUES ADOPTÉES POUR LA CRÉATION D'UN CORPUS THÉMATIQUE DANS LE CADRE DU PROJET OURAL

Nathalie GASIGLIA
UMR Savoirs, Textes, Langage (UMR 8163 du CNRS) - Université Lille 3

1. CONTEXTE DE TRAVAIL : PRÉSENTATION DU PROJET OURAL

En 2002 débutait l'action Technolangue[1]. À cette époque, Claude de Loupy, responsable recherche chez Sinequa S.A.S., intervenait dans le cadre de la formation lilloise de lexicographes et terminographes[2], où il contribuait à faire réfléchir les étudiants sur les qualités particulières des descriptions lexicales dédiées à des applications talistes. Pour pousser plus loin différentes collaborations engagées, dont la nôtre, Claude de Loupy s'est entouré de cinq partenaires[3] et a introduit une demande de financement. Le nom du projet que nous avons élaboré – OuRAL pour **Ou**tils et **R**essources pour l'**A**nalyse de la **L**angue – témoigne de la diversité des interactions qui nous réunissaient par sous-groupes et qui nous ont fait nous engager à développer d'une part un lexique et des corpus, et d'autre part des outils d'étiquetage morphosyntaxique, de repérage d'entités et de segmentation en vue de la constitution de résumés. L'UMR SILEX[4], où je coordonnais le travail, devait œuvrer avec Sinequa et le VALORIA à la constitution de corpus constitués de textes écrits et de transcriptions de dialogues, structurés en XML et étiquetés morphosyntaxiquement et au niveau de certaines entités. Je ne me concentrerai plus ci-après que sur les transcriptions et annotations réalisées dans mon UMR.

Le projet AGILE, coordonné par Bernard Normier, a réuni OuRAL et d'autres projets[5].

2. PRÉSENTATION DU SOUS-PROJET LILLOIS

2.1. Concilier les intérêts locaux pour l'étude lexicale et les intérêts des partenaires

Quand nous avons envisagé de développer un corpus d'oral transcrit, le VALORIA était déjà engagé dans des travaux de ce type – il avait constitué deux corpus de dialogue oral, OTG et ECOLE_MASSY – et cherchait à donner une visibilité et une portée plus larges aux travaux développés en ingénierie des langues, appliquée au monde du handicap. Il semblait efficace, pour une collaboration effective entre les partenaires du projet OuRAL, que l'effort entrepris à l'UMR SILEX soit orienté vers la création d'un corpus de dialogues radiodiffusés relatifs à des thèmes particuliers – conformément aux hypothèses de haute rentabilité d'exploration pour documenter l'analyse lexicologique que j'avais formulées (Gasiglia 2004 et 2005) –, et que nous sélectionnions des échanges relatifs à des domaines

qui concernent les personnes handicapées auxquelles les recherches du VALORIA sont destinées.

Le travail de constitution du corpus de dialogues transcrits a débuté par une étude prospective des émissions radiodiffusées thématiques présentant des tours de parole riches en contenus informationnels et l'emploi d'un vocabulaire précis, donnant éventuellement lieu à des reformulations ou des explicitations. Deux émissions ont été retenues. Elles abordent différents thèmes de la vie quotidienne et font intervenir des auditeurs, des spécialistes d'un domaine (juristes ou techniciens) et un animateur. Une sélection des émissions diffusées durant les deuxième et troisième trimestres 2003 privilégiait celles relatives à des questions juridiques induites par un handicap physique ou mental, ou à des problèmes techniques présentant un intérêt particulier dans le cadre de l'accompagnement des personnes handicapées. Le travail de transcription (fait avec le logiciel Transcriber version 1.4.2[6], cf. § 3.2) a été réalisé pour neuf émissions, soit trois heures environ d'échanges (hors publicités et annonces diverses), ce qui représente un volume de plus de 55 000 mots.

2.2. Inflexion pour cause de non-obtention des droits

Le corpus développé dans le cadre du projet OuRAL doit être librement diffusable. Afin de garantir cette diffusabilité, dès le début de la sélection des émissions, des contacts ont été pris avec les chaînes de radio afin d'entrer en relation avec les orateurs intervenant dans les émissions retenues pour leur demander le droit de transcrire leurs paroles et de diffuser le corpus ainsi constitué. Mais fin 2003 les démarches entreprises n'avaient pas abouti, personne ne semblant avoir le pouvoir de négocier ces questions dans les grandes radios concernées. Sans autorisations obtenues avant 2005, le produit du travail engagé n'aurait pas été livrable dans le cadre du projet OuRAL : il était impératif de réviser notre projet.

Partant, en décembre 2003, un thème unique a été choisi : la gestion locative de biens immobiliers, thème qui touche une large part de la population, qui dispose d'un vocabulaire spécifique mais employé avec une assurance inégale par les protagonistes des négociations et démarches, et qui doit pouvoir apporter des éléments utiles pour des applications talistes et plus particulièrement pour l'élaboration d'outils d'aide à la communication. Une campagne d'entretiens a été lancée. Faute de moyens[7], ils ne pouvaient pas faire intervenir un animateur, des néophytes et des spécialistes comme dans les émissions radiophoniques retenues, mais ils impliquaient des locataires, des propriétaires bailleurs et des personnels d'agences immobilières[8], à qui nous faisions signer une autorisation d'exploitation de leurs propos dans le cadre du projet OuRAL. Quatorze entretiens de vingt-cinq à quatre-vingt-dix minutes ont été enregistrés et transcrits, ce qui correspond à 99 000 mots environ.

3. DÉROULEMENT TECHNIQUE DU TRAVAIL ÉFFECTUÉ A L'UMR SILEX

3.1. Cadrage global : de l'organisation des entretiens à l'étiquetage des transcriptions

Les entretiens ont été organisés durant les deux premiers trimestres 2004 et ont mobilisé des membres du projet et des volontaires recrutés par annonce et sélectionnés selon des critères nous permettant d'équilibrer les types d'expériences (locataire, propriétaire bailleur

ou chargé de gestion locative) et les profils sociaux (étudiant ou non, salarié ou non, de nationalité correspondant au pays où le bien était loué ou non, vivant seul ou non, etc.).

Toutes les transcriptions ont été effectuées sous Transcriber (version 1.4.2, cf. § 3.2). Après la seconde relecture, le nombre d'infidélités de transcription et de fautes d'orthographe était encore non négligeable, mais les contraintes de calendrier et de personnel ont interdit d'envisager de nouvelles relectures avant l'étiquetage grammatical. Nous avons donc dû considérer que les problèmes restants seraient corrigés lors du contrôle des lemmes et étiquettes morphosyntaxiques.

Les textes des transcriptions ont ensuite été traités par Cordial Analyseur (version 7.0). L'utilisation de cet étiqueteur a été décidée après évaluation de ses performances sur ce type particulier de données (cf. § 3.3). Cet outil n'étant pas en mesure d'enjamber les balises XML contenues dans les transcriptions alignées avec leurs sources sonores, les documents produits avec Transcriber ont dû être prétraités (cf. § 3.4). Deux analyses ont été demandées à Cordial Analyseur, l'une avec reconnaissance de toutes les expressions[9], l'autre sans. Deux versions étiquetées ont donc été obtenues pour chaque transcription et ont ensuite été fusionnées puis converties en XML, avec mention des deux niveaux d'analyse pour les segments qui peuvent être considérés comme étant des expressions – au niveau des mots et au niveau des expressions (cf. § 3.5). Les étiquetages produits par Cordial Analyseur, synthétisés et XMLisés, ont bénéficié d'une relecture minutieuse. Compte tenu de la difficulté éprouvée à proposer des étiquetages fiables dans certains contextes, nous avons décidé d'introduire des commentaires ponctuels, qui précisent les recherches faites dans des dictionnaires et des ouvrages de linguistique strictement référencés ou l'état de la réflexion des relecteurs, et d'enrichir le jeu d'étiquettes morphosyntaxiques employé (cf. § 3.6, et Mostrov, à paraître). L'ensemble des transcriptions a enfin été anonymisé (cf. § 3.8) – ce qui n'a pas été fait avant, pour ne pas perturber l'étiquetage et les relectures –, mais aucune des entités qui devaient y être balisées n'a finalement pu être repérée (cf. § 3.7).

3.2. Transcription des entretiens avec Transcriber 1.4.2

Pour les entretiens comme pour les émissions radiophoniques traitées initialement dans le cadre de ce projet ou pour la constitution du *Corpus foot de multiplex transcrits*[10], les transcriptions ont été réalisées avec Transcriber 1.4.2. Les documents sonores ont été segmentés en tours de parole, auxquels sont associées des informations sur la ou les personnes qui s'expriment et où sont repérés divers événements linguistiques comme des troncations de mots, des prononciations remarquables, etc. (cf. Gasiglia 2006 : § 2.2). Transcriber 1.4.2 prévoit dans son interface la saisie d'une large partie de ces informations :

• pour la caractérisation des locuteurs, on peut aisément donner leur nom, spécifier leur sexe ou indiquer s'ils semblent être francophones natifs, mais la fenêtre de saisie ne prévoit aucun repérage sociolinguistique et elle ne permet pas de noter les valeurs des différents critères de sélection concernant nos intervenants (cf. § 3.1) ;

• pour les annotations internes à chaque tour de parole, des événements sont prédéfinis, d'autres sont définissables, et des commentaires peuvent être introduits. Ces deux types d'éléments ont permis de coder tous les objets que nous souhaitions repérer[11].

3.3. Choix de Cordial Analyseur 7.0

Les textes des transcriptions ont été traités par Cordial Analyseur 7.0, mobilisé du fait du retard de développement de l'étiqueteur morphosyntaxique en chantier dans le cadre du projet OuRAL. Deux raisons majeures ont motivé le choix de Cordial Analyseur : d'une part cet étiqueteur était déjà exploité au VALORIA pour traiter des corpus d'oral transcrit, d'autre part nous l'utilisions nous-mêmes régulièrement.[12] Cependant, en dehors des problèmes de format de fichiers, sur lesquels je reviendrai, deux points nourrissaient nos inquiétudes relativement à son emploi pour traiter les transcriptions issues de Transcriber.

• Le premier est technique : nous voulions demander à Cordial Analyseur d'analyser des tours de parole non ponctués et parfois aussi riches en ruptures que longs, alors que nous savions que son fonctionnement est indexé sur la forme normalisée des documents écrits et qu'il ne sait pas analyser de séquence de plus de 127 mots. Par ailleurs, dans les dialogues transcrits à étiqueter, il pouvait y avoir des recouvrements de paroles, un second locuteur commençant à parler alors que le premier n'a pas terminé son énoncé, ce qui l'amène soit à s'interrompre (en laissant éventuellement une structure argumentale incomplète), soit à réorienter son propos. Pour les analyses linguistiques utilisant ces transcriptions, il serait plus pertinent de disposer de chaque énoncé dans sa globalité, avec, au sein de chacun, un marquage des zones de recouvrement de paroles. Ce découpage serait également mieux adapté pour l'analyse syntaxique que Cordial Analyseur effectue, mais Transcriber ne sait pas associer un même segment sonore à deux sous-parties de deux tours de parole distincts : un nouveau tour de parole doit en conséquence être créé pour chaque zone de recouvrement avec, avant et après, des tours de parole pour les parties d'énoncés prononcées par un seul locuteur. Différents tests ont été effectués pour évaluer si nous devions mettre en place des prétraitements du corpus, voire réviser nos options de transcription, afin que l'intégralité de chaque énoncé soit prise en considération lors de l'étiquetage malgré (i) l'absence de ponctuation interne des phrases, (ii) la segmentation des énoncés en tronçons de 127 mots et (iii) la répartition dans deux tours de parole des propos énoncés par chaque orateur en cas de chevauchements. Dans la mesure où les analyses demandées à Cordial Analyseur sont circonscrites aux lemmes et aux codes morphosyntaxiques, ces contraintes n'induisent des erreurs qu'en présence de graphies correspondant à plusieurs unités de catégories différentes, employées dans des contextes rendus ambigus par ces segmentations, ce qui s'est révélé rare. Aucun prétraitement n'a donc été nécessaire, mais les relecteurs ont dû prêter une attention particulière aux zones à risques.

• Le second point est une question de fond : un projet comme OuRAL peut-il proposer un format d'étiquetage propre à un outil ou doit-il être aussi normalisé que possible ? Les partenaires du projet OuRAL n'ont ni été en mesure de décider à quel standard d'étiquetage morphosyntaxique se conformer ni su définir une manière commune. L'emploi du même étiqueteur pour le corpus de textes et les corpus d'oral transcrit aurait harmonisé les productions, à défaut de fonder un choix méthodologique, mais, l'étiqueteur en chantier dans le cadre d'OuRAL n'étant pas prêt, nous devions utiliser d'autres outils (Cordial Analyseur au VALORIA et chez nous, un outil interne chez Sinequa) puis envisager de recoder les étiquetages pour les harmoniser. Ce travail semblait *a priori* lourd dans la mesure où il ne se limitait pas à convertir les noms et les valeurs des attributs (cf. § 3.5) : en fonction des principes d'analyse linguistique adoptés par chaque étiqueteur, les codes morphosyntaxiques ne sont pas en relation bijective, ce qui fait que les valeurs ne peuvent pas être transcodées automatiquement. Afin d'évaluer la faisabilité de cette

conversion, dès 2003, avec deux stagiaires, j'avais comparé les jeux d'étiquettes employés par les différents partenaires et fait des projections pour élaborer un jeu unique. Ce travail avait donné lieu à un rapport mais il présentait un vice de fond qui en biaisait les conclusions : nous ne disposions pas d'un corpus étiqueté par chaque système, mais de listes d'étiquettes associées à leur forme développée. Nous ne pouvions donc pas comparer les étiquetages ni les valeurs relatives de chaque code. Finalement, nous n'avons pas recodé les étiquetages des transcriptions.

3.4. Prétraitement des sorties de Transcriber en vue de leur manipulation par Cordial Analyseur

Transcriber produit des documents XML où les données transcrites sont balisées et donc les caractères du balisage mêlés à ceux des transcriptions. Cordial Analyseur n'est pas conçu pour analyser des documents XML[13], il ne sait pas enjamber les balises[14] et va traiter ce qui y est inclus au même titre que les données, comme le montre la Figure 1, où le contenu de la balise ouvrante d'un élément *Turn* est analysé, dans la colonne de gauche, comme celui de la balise fermante après les mots de l'énoncé transcrit dans la colonne de droite.

\r<			\r		
Turn	Turn	NPI	euh	euh	INT
speaker	speaker	NCMS	au niveau des	au niveau des	PREP
="			garants	garant	NCMP
AJ	AJ	NPFS	donc	donc	COO
"			euh	euh	INT
startTime="108.0	starttime-108.0	NCI	y	y	PPER3S
"			a	avoir	VINDP3S
endTime="154.459	endtime-154.459	NCI	quatre	quatre	ADJNUM
">			garants	garant	NCMP
\r<			[…]		
Sync	Sync	NPI	\r</		
time="103.498	time-103.498	NCI	Turn	Turn	NPI
"/>			>		

Figure 1 : Tour de parole transcrit avec Transcriber et étiqueté avec Cordial Analyseur

Il s'avérait donc nécessaire d'intervenir sur le balisage des fichiers produits par Transcriber pour le rendre plus concis (sans perdre d'information non restaurable) et de mettre en place des scripts de nettoyage à appliquer aux fichiers produits par Cordial Analyseur pour effacer les informations parasites associées au balisage des transcriptions. La réflexion relative à la première de ces deux étapes a été menée en tenant compte de la DTD qui avait été utilisée pour le premier *Corpus foot de multiplex transcrits* (cf. note 10), de manière à ce que l'ensemble des corpus développés sous ma direction soit cohérent et que leurs données puissent être manipulées par les mêmes scripts Perl ou routines XSLT[15].

Les fichiers obtenus après conversion du balisage, étiquetage par Cordial Analyseur et nettoyage contiennent des triplets mot, lemme, code morphosyntaxique (cf. Figure 2).

<SPEAKER IDENTITY="AJ"			y	y	PPER3S
POST="locataire" startTime="138.0">			a	avoir	VINDP3S
euh	euh	INT	quatre	quatre	ADJNUM
au niveau des	au niveau des	PREP	garants	garant	NCMP
garants	garant	NCMP	[…]		
donc	donc	COO	</SPEAKER>		
euh	euh	INT			

Figure 2 : Tour de parole transcrit, converti, étiqueté avec Cordial Analyseur et nettoyé

3.5. Fusion et XMLisation des produits de Cordial Analyseur

Chaque transcription a donné lieu à deux versions d'étiquetage, l'une avec repérage global de toutes les expressions que Cordial Analyseur sait reconnaître (cf. note 9 et Figures 1 et 2), l'autre avec analyse de tous les items isolément. Ces deux versions ont été comparées (avec la commande `sdiff` du shell Linux) et fusionnées afin de proposer, chaque fois que cela semblait pertinent d'un point de vue linguistique, un étiquetage des expressions considérées dans leur globalité et de chaque item constituant (cf. Figure 3).

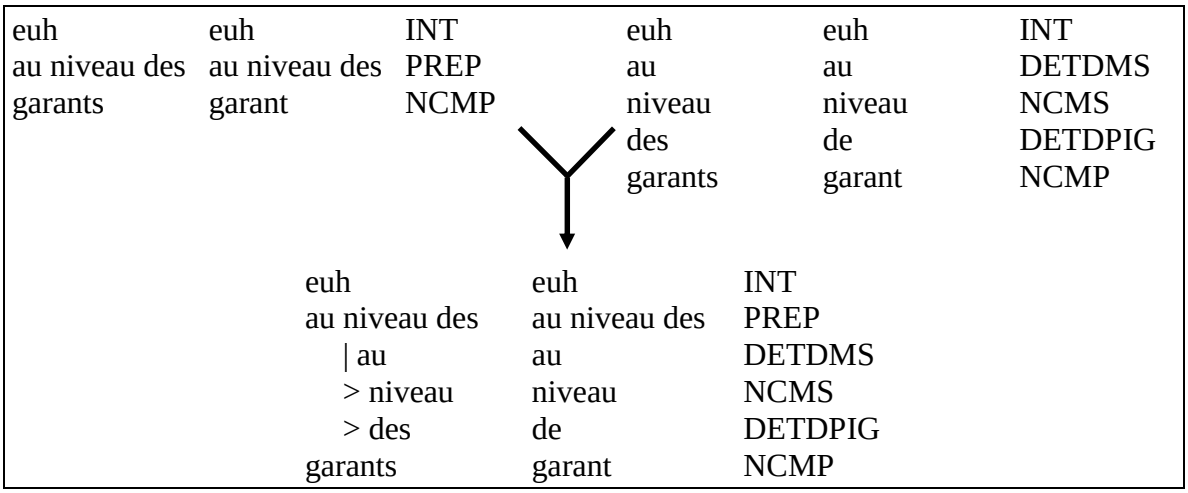


Figure 3 : Fusion des étiquetages avec et sans repérage d'expressions

Les étiquetages fusionnés ont enfin été convertis en XML, en associant deux attributs à chaque unité (*ITEM* ou *EXPRESSION*) : l'un pour le lemme, l'autre pour le code morphosyntaxique (cf. Figure 4). Ce balisage « maison » ne devait être que provisoire. Il devait être converti une fois qu'une norme d'étiquetage aurait été adoptée (cf. § 3.3).

```
<ITEM lemme="euh" typegram="INT">euh</ITEM>
<EXPRESSION lemme="au niveau des" typegram="PREP">
    <ITEM lemme="au" typegram="DETDMS">au</ITEM>
    <ITEM lemme="niveau" typegram="NCMS">niveau</ITEM>
    <ITEM lemme="de" typegram="DETDPIG">des</ITEM>
</EXPRESSION>
<ITEM lemme="garant" typegram="NCMP">garants</ITEM>
```

Figure 4 : Conversion en XML de l'étiquetage fusionné

3.6. Relecture des transcriptions étiquetées et enrichissement du jeu d'étiquettes

La relecture des transcriptions étiquetées s'est révélée beaucoup plus lente que nous l'avions envisagé et ceci tient à deux facteurs, d'une part le manque d'expérience en la matière qui nous avait certainement conduits à sous-estimer le coût de cette tâche minutieuse ; d'autre part, la quantité des corrections et la difficulté à proposer des étiquetages fiables dans certains contextes. Compte tenu de l'ampleur des problèmes rencontrés, nous avons décidé d'enrichir les documents relus de commentaires, précisant (i) les recherches faites sur tel ou tel point dans des dictionnaires et des ouvrages de linguistique strictement référencés, et (ii) l'état de la réflexion des relecteurs sur des points jugés problématiques. Dans la majorité des cas, ces commentaires pointent des hésitations ou des ambiguïtés locales, comme celle présente dans l'extrait ci-dessous :

> *le propriétaire a le droit de rompre le bail… il a le droit tout le temps mais je crois qu'**il** a des différences de motifs et de délais*

où il est malaisé de trancher entre deux analyses pour la seconde occurrence du pronom ***il*** : s'agit-il d'une nouvelle référence au propriétaire, ce qui induira un étiquetage comme pronom personnel 3e personne masculin singulier ou comme pronom impersonnel si, comme il est courant dans la région du Nord, *il a* s'interprète comme *il y a* ?

Les conditions dans lesquelles ce travail a été fait ont permis à Vassil Mostrov, qui a assumé une part importante des relectures, de mener une réflexion sur le degré de raffinement souhaitable pour ce type d'étiquetage et d'enrichir le jeu d'étiquettes de Cordial Analyseur, puis d'évaluer son applicabilité et sa pertinence sur le corpus (cf. Mostrov, à paraître). Ici, son travail conduira à choisir entre les codes *PPER3MS* et *PPER3N*, selon l'analyse retenue, au lieu de l'étiquette unique *PPER3S* proposée par Cordial Analyseur.

3.7. Balisage d'entités

Dans le texte présentant les résultats attendus à la fin du projet OuRAL, nous avions indiqué qu'à l'étiquetage morphosyntaxique s'ajouterait le balisage d'entités caractéristiques comme les dates, les noms de personnes, les lieux, etc., mais la liste précise de ces entités restait à déterminer. Peu d'entités du type de celles évoquées ci-dessus sont présentes dans le corpus de gestion locative. À peine retrouvons-nous des noms de villes ou de rues, à raison d'une dizaine d'occurrences par entretien. Les partenaires n'ayant pas élaboré de listage précis des entités à repérer ni de codification de leur balisage, nous avons pris le parti de ne pas remplacer les noms de villes ou de rues présents dans les transcriptions lors de leur anonymisation, et ainsi de permettre leur repérage ultérieur.

3.8. Anonymisation des transcriptions

Afin de respecter l'anonymat des personnes impliquées dans ces entretiens et de garantir autant que possible que les orateurs ne seront pas identifiables, leur identité, déclinée en début d'entretien, est gommée, et les numéros de rue modifiés dans les mentions d'adresses. En outre, lors de notre engagement dans le projet OuRAL, nous pensions travailler sur des sources radiodiffusées (cf. § 2), et nous ne nous étions pas engagés à fournir les enregistrements sonores des émissions transcrites. Bien que nous ayons changé de sources, nous avons décidé de ne pas livrer les enregistrements, afin de garantir

l'anonymat des personnes dont la voix pourrait être reconnue, et de ne fournir que des transcriptions non alignées.

3.9. Corpus livré dans le cadre du projet

Seule une partie des transcriptions réalisées durant les deux années de travail sera diffusable, faute d'avoir obtenu les droits concernant les émissions radiodiffusées (cf. § 2). Le corpus livré à la fin du projet est constitué des entretiens relatifs à la gestion locative de biens immobiliers faisant intervenir des locataires, des propriétaires bailleurs et des animateurs/intervieweurs. C'est une version XML étiquetée et relue des transcriptions qui est livrée, une version où tous les items portent un code morphosyntaxique et la mention de leur lemme, où les expressions sont analysées globalement et décomposées, et où les étiquetages les moins assurés sont accompagnés d'un commentaire motivant les décisions prises. Il s'agit donc d'un corpus d'oral transcrit enrichi et révisé d'un peu moins de 100 000 mots.

4. ÉVALUATION QUALITATIVE DU PRODUIT LIVRÉ

En 2003, aux Journées de linguistique de corpus de Lorient, et dans des articles publiés en 2004 et 2005, j'avais motivé la création d'un corpus de dialogues radiodiffusés relatifs à des thèmes particuliers en formulant des hypothèses de haute rentabilité d'exploration pour documenter l'analyse lexicologique. Le travail entrepris avec Pierre Corbin et les étudiants de la formation LTTAC (cf. note 2) a donné de premiers résultats encourageants concernant un thème sportif (cf. note 10), et la participation à OuRAL devait me donner les moyens de développer d'autres transcriptions relatives à d'autres thèmes, afin de tester sur d'autres terrains la validité de mes hypothèses. En cherchant à sélectionner des échanges verbaux intervenant dans des domaines de la vie courante qui pouvaient concerner les personnes handicapées ciblées par les recherches du VALORIA, j'ai cherché à doublement motiver le travail dans lequel je me suis engagée et auquel j'ai associé quelques jeunes chercheurs.

Le corpus relatif à la gestion locative de biens immobiliers qui a finalement été développé n'a pas encore été exploré de manière systématique, et sa qualité relative pour la documentation de descriptions lexicales reste à apprécier. Mais au cours des relectures des transcriptions j'ai été sensible à deux tendances :

(i) les dialogues de nos entretiens sont constitués de prises de parole beaucoup plus fractionnées que celles des émissions radiodiffusées, avec un rythme d'échanges très rapide du fait de la relative brièveté des développements fournis en guise de réponses, que cette brièveté soit due à la précision même des questions qui n'appellent qu'un mot de réponse, au manque de recul des locuteurs face aux démarches qu'ils ont accomplies ou à d'autres paramètres que je n'ai pas encore identifiés ;

(ii) la richesse lexicale y est peut-être relativement faible, puisque si nous retrouvons bien dans le corpus une large part du vocabulaire propre à la gestion locative de biens immobiliers (*bail*, *état des lieux*, *garant*, etc.), ces items n'apparaissent éventuellement pas dans des constructions typiques que je connais ni dans des patrons syntaxiques qui frapperaient par leur régularité.

Ce corpus étant livré en fin de projet, il pourra alimenter d'autres recherches. Peut-être pour elles ne présentera-t-il pas les faiblesses que je crains d'y discerner dans le cadre de

l'analyse lexicale – probablement en présentera-t-il d'autres. Je souhaite en tout cas qu'il puisse constituer une ressource de qualité apte à nourrir le travail d'autres chercheurs.

5. REMERCIEMENTS

Ce travail n'aurait pas eu lieu sans l'impulsion donnée par l'action Technolangue ni les souhaits de collaboration plus poussée avec Claude de Loupy et quelques autres partenaires du projet OuRAL. Ce travail n'aurait pas pu être mené à son terme sans le travail assidu de Vassil Mostrov et des autres stagiaires et vacataires : Bichara Assoumani et Sibel Maras (en 2003), puis Olivier Cavallaro et Jang Martel (en 2004), et enfin Gergana Shtereva (en 2005), tous issus des formations talistes lilloises et avec qui nous avons partagé des interactions stimulantes. Il n'aurait pas non plus pu être accompli sans le prêt de matériel informatique par l'entreprise Sinequa et sa gestion par l'administrateur système de notre UMR, ni sans la collaboration d'un certain nombre de personnels administratifs, à l'UMR, au service de la valorisation de la recherche et à la DRH de l'université Lille 3. Il n'aurait enfin pas pu être engagé sans la générosité des locuteurs anonymes dont nous avons enregistré les dialogues, qui nous ont offert leur temps et leur expérience et qui nous ont accordé le droit d'exploiter les transcriptions de leurs propos.

Ce texte pour sa part doit beaucoup à Geoffrey Williams et Christophe Ropers, qui ont organisé les quatrièmes journées de linguistique de corpus à Lorient en 2005 et qui nous ont ensuite offert un espace de publication. Il a aussi et surtout bénéficié des critiques et de relectures minutieuses de Pierre Corbin et de Vassil Mostrov.

Merci à tous.

6. NOTES

[1] L'action Technolangue a été conduite par le Ministère de la Recherche et des Nouvelles Technologies, en partenariat avec le Ministère délégué à l'Industrie et le Ministère de la Culture et de la Communication, cf. http://www.technolangue.net/rubrique22.html.

[2] Le DESS LTTAC, puis Master 2e année "Lexicographie, Terminographie et Traitement Automatique des Corpus", formation professionnalisante de la spécialité "TAL, Dictionnaires, Terminologies, Corpus" du master "Arts, Lettres, Langues et Communication" mention "Sciences du langage".

[3] Le Laboratoire de Psychologie Expérimentale (UMR 8581 - Université Paris V), le Laboratoire d'Informatique de Paris 6, le Laboratoire Informatique d'Avignon, le Valoria (EA 2593 - Université de Bretagne-Sud) et l'UMR SILEX maintenant UMR STL, cf. note 4.

[4] En 2002, quand l'action Technolangue a débuté, l'UMR SILEX (UMR 8528 du CNRS) n'avait pas encore fusionné avec l'UMR Savoirs et Textes. Cette fusion est effective depuis janvier 2006, elle a donné naissance à l'UMR Savoirs, Textes, Langage (STL). Le travail réalisé dans le cadre de l'action Technolangue l'ayant été du temps de l'UMR SILEX, je ne mentionnerai plus ici que ce nom.

[5] D'autres projets, conçus de manière indépendante, ont étés réunis au sein du projet AGILE : ALIZE (coordinateur : Ivan Magrin-Chagnolleau (CNRS)), TILT (coordinatrice : Nicole West (AFNOR)) et WATSON (coordinateur : José Coch (Lingway)).

[6] Transcriber est un éditeur XML de transcriptions alignées au son capable de traiter des fichiers son de taille importante, cf. http://sourceforge.net/projects/trans/ et Barras *& al.* (2000 et 2001), Geoffrois *& al.* (2000).

[7] Aucun budget n'avait été prévu pour le paiement d'honoraires de juristes ou de techniciens ou pour le défraiement des personnes participant à ces entretiens.

[8] Finalement aucun personnel d'agence immobilière n'a pu participer aux entretiens.

[9] Dans la fenêtre de paramétrage des étiquetages morphosyntaxiques, nous avons demandé que soient reconnues toutes les expressions : les locutions prépositionnelles, conjonctives, pronominales et adjectivales indéfinies, et les expressions adverbiales, adjectivales, nominales et verbales.

[10] Le *Corpus foot de multiplex transcrits* est un corpus xmlisé développé depuis 2002 avec Pierre Corbin et les étudiants de la formation LTTAC (cf. note 2). Les premières transcriptions n'étaient pas alignées avec les

commentaires de matchs de football enregistrés et avaient demandé l'élaboration d'une DTD où étaient déclarés divers éléments qui nous permettaient de repérer ce que nous jugions utile de pointer pour les études lexicales auxquelles les transcriptions devaient servir (cf. Gasiglia 2004 et 2006 : § 2.2). Une seconde vague de transcriptions de commentaires de matchs de football a été réalisée sous Transcriber (Gasiglia 2006 : § 2.3).

[11] Dans la version 1.5.1 de Transcriber rendue disponible en 2005, les repérages d'entités nommées sont prévus ainsi que diverses autres annotations, ce qui aurait allégé le balisage mis en œuvre. Mais les transcriptions d'entretiens étaient déjà réalisées au moment de la mise en ligne de cette nouvelle version et, compte tenu de la date d'échéance du projet OuRAL, elles n'ont pas été révisées en exploitant les nouvelles fonctionnalités disponibles.

[12] Notons que Cordial Analyseur a par ailleurs été exploité pour étiqueter les textes de la FreeBank, les étiquettes étant ensuite converties en annotations XML externes (Salmon-Alt & *al.* 2004 : 158).

[13] Les formats de fichiers manipulables par Cordial Analyseur, outre ceux qu'il génère, correspondent aux extensions .txt, .rtf, .htm, .doc (Word2 à Word6), .wp (WordPerfect), .sam (AmiPro).

[14] La version 8.13 de Cordial Analyseur a été testée sans qu'aucun progrès soit observé sur ce point.

[15] La mise à disposition de nouvelles fonctionnalités dans les versions récentes de Transcriber conjuguée au souhait d'aligner les premières transcriptions du corpus footballistique au son des enregistrements des multiplex m'amènera à réviser le balisage mis en œuvre pour le rendre conforme à la DTD de Transcriber, mais il était impossible de mener ce chantier à bien durant le déroulement du projet OuRAL. Il est donc apparu plus réaliste, lors de la manipulation du balisage des transcriptions d'entretiens pour leur étiquetage, de travailler avec le jeu de balises antérieurement stabilisé (cf. Gasiglia 2006 : § 2.2).

7. REFERENCES

Barras C., Geoffrois E., Wu Z. & Liberman M. 2000. 'Transcriber: a free tool for segmenting, labelling and transcribing speech'. Actes de la conférence LREC (Second International Conference on Language Resources and Evaluation), p. 1373-1376.

Barras C., Geoffrois E., Wu Z. & Liberman M. 2001. 'Transcriber: development and use of a tool for assisting speech corpora production'. Speech Communication (special issue on Speech Annotation and Corpus Tools), vol. 33, No 1-2, p. 5-22.

Gasiglia N. 2004. 'Faire coopérer deux concordanciers-analyseurs pour optimiser les extractions en corpus' dans Habert B. (dir.) 2004. *Linguistique et informatique : nouveaux défis*. Revue Française de Linguistique Appliquée, vol. IX-1, p. 45-62.

Gasiglia N. 2005. 'Stratégie de constitution de corpus oraux transcrits : arguments pour un corpus plurithématique à haut rendement' dans Williams G. (éd.) 2005. *La linguistique de corpus en France ou en français*. Presses Universitaires de Rennes, p. 219-232.

Gasiglia N. 2006. 'Stratégie de consultation de corpus oraux transcrits : pistes méthodologiques pour l'exploration d'un corpus thématique à haut rendement' dans Williams G. (éd.) 2006. *Texte et Corpus*. Presses Universitaires de Rennes.

Geoffrois E., Barras C., Bird S. & Wu Z. 2000. 'Transcribing with Annotation Graphs'. Actes de la conférence LREC (Second International Conference on Language Resources and Evaluation), p. 1517-1521.

Habert B. 2005. *Instruments et ressources électroniques pour le Français*. Paris : Ophrys.

Habert B., Fabre C. & Issac F. 1998. *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*. Paris : InterEdition.

Ide N. & Véronis J. 1996. 'Une application de la TEI aux industries de la langue : le Corpus Encoding Standard'. Cahiers GUTenberg n° 24 (spécial TEI), p. 166-169.

Mostrov V. à paraître. 'Appréciations sur la fiabilité d'un étiquetage morphosyntaxique généré automatiquement et présentation du travail lié à son enrichissement' dans Williams G. (éd.) (à paraître), *Actes des Quatrièmes journées de linguistique de corpus*. Presses Universitaires de Rennes.

Salmon-Alt S., Romary L. & Pierrel J.-M. 2004. ‘Un modèle générique d’organisation de corpus en ligne’ dans Blache P., Bonastre J.-F. & Nguyen N. (dir.) 2004. *Traitement de l’écrit et de la parole*. Revue TAL, vol. 45, n° 3, p. 145-169.

Valli A. & Véronis J. 1999. ‘Étiquetage grammatical des corpus de parole : problèmes et perspectives’ dans *L’oral spontané*. Revue Française de Linguistique Appliquée, vol. IV-2, p. 113-133.

Véronis J. 2000. ‘Annotation automatique de corpus : panorama et état de la technique’ dans Pierrel J.-M. (éd.) 2000. *Ingénierie des langues*. Paris : Hermès, p. 111-129.

TRANSCRIPTION PROSODIQUE NORMALISÉE AU SEIN DU PROJET PFC (PHONOLOGIE DU FRANÇAIS CONTEMPORAIN) : L'ÉTAT D'UN CHANTIER

Anne Lacheret-Dujour[1], Chantal Lyche[2]
[1]Institut Universitaire de France, Paris,
Laboratoire CRISCO, Université de Caen 14032 Caen Cedex
[2]Universitetet i Oslo/Tromsø
Anne.lacheret@crisco.unicaen.fr ; chantal.lyche@ilos.uio.no

1. INTRODUCTION

L'objectif de notre communication est d'exposer l'approche retenue pour le codage prosodique de la base de données PFC (Phonologie du Français Contemporain), amorcé en décembre 2003. Après avoir présenté le projet dans ses grandes lignes, et défini le domaine de la prosodie sous l'angle phonétique, phonologique et symbolique, il s'agira de discuter les limites tant théoriques que pratiques de ce codage, d'expliquer pourquoi les secondes découlent en partie des premières, illustrant ainsi le fait qu'il n'existe pas d'outils et d'instrumentation utilisables et robustes sans un minimum épistémologique commun, cela quel que soit le niveau de traitement linguistique concerné, en l'occurrence le niveau suprasegmental. En pratique, nous essaierons de comprendre pourquoi ce minimum commun est particulièrement difficile à stabiliser quand il s'agit de travailler sur des données prosodiques, données effectivement particulièrement résistantes à l'approche structuraliste du linguiste d'une part, à l'instrumentation d'autre part, enfin au traitement automatique. Dans ce contexte, nous présenterons les objectifs actuels du codage prosodique dans PFC, discuterons des logiciels utilisables pour ce codage et pour finir notre méthodologie de transcription sera présentée. Pour clore cette présentation, nous conclurons sur nos perspectives de recherche après avoir résumé les hypothèses fortes qui sous-tendent notre démarche.

2. LE PROJET PFC

Le propos de ce paragraphe est de rappeler les objectifs du projet PFC (voir aussi Durand & Lyche, 2003) et le domaine de la prosodie sous l'angle substantiel (phonétique), formel (phonologique) et fonctionnel (fonctions syntaxiques et pragmatiques de l'accent et de l'intonation).

2.1. Les objectifs du projet

Le projet *Phonologie du français contemporain : usages, variétés et structures* (http://www.projet-pfc.net) se situe dans la tradition des grandes enquêtes phonologiques et vise à constituer sur 50 points d'enquête un grand corpus oral d'environ 500 locuteurs, constitué de données strictement comparables. Tous les enregistrements sont effectués sur la base d'un protocole unique avec la prise en compte de quatre registres (liste de mots et

texte lu, conversations dirigée et libre). L'objectif premier de PFC est de nature phonologique. Le corpus constitué doit permettre des analyses de phénomènes typiquement variables, et ce faisant de tester les modèles phonologiques et phonétiques qui accordent une large place à la variation. Dans un premier temps l'accent est mis sur le système phonémique du locuteur, sur la prononciation variable du « e » caduc (schwa) et de la liaison. Pour chaque locuteur, le texte ainsi que quelques minutes d'entretien sont codés sous PRAAT, sur deux tires traitant respectivement le schwa et la liaison. Le codage, effectué sur la base de la transcription orthographique, est alphanumérique ; son rôle est d'offrir un premier balayage des données pour les analyses ultérieures. Des principes de codage équivalents ont été envisagés pour la transcription de la prosodie tout en sachant que celle-ci, très lourde à mettre en œuvre et à exploiter réclame également ses propres contraintes de codage (cf. *infra*, 3.3).

2.2. Pourquoi la prosodie dans le projet PFC ?

D'une manière générale, la transcription prosodique poursuit deux pistes dont seule la première concerne les auteurs du papier : (i) mettre en valeur le lien entre deux phénomènes hautement variationnels (le schwa et la liaison) et la prosodie, et (ii) dégager les caractéristiques prosodiques générales de chaque variété de français étudiée dans un but purement descriptif et comparatif. Dans cette communication, nous nous centrerons exclusivement sur l'interaction schwa/prosodie.

2.3. Le domaine de la prosodie : substance, forme, fonction

Tenter de saisir la portée définitoire du domaine n'est pas une mince affaire (Lacheret & Beaugendre, 1999), tant les définitions sont multiples, plus ou moins opaques et circulaires, centrées sur différents niveaux d'analyse maladroitement amalgamés : paramètres phonétiques actualisés dans la substance, formes prosodiques rattachées au système de la langue, fonctions linguistiques et expressive de la prosodie. Dans ce contexte, il nous semble raisonnable d'envisager une définition modulaire du domaine, de ses caractéristiques substantielles aux fonctions que la prosodie assume dans le langage, même si en pratique, les différents niveaux de traitement interagissent étroitement. La figure 1 ci dessous présente pour chaque niveau (phonétique vs. phonologique) les objets prosodiques manipulés ; les événements phonétiques (en premier lieu fréquence fondamentale et durée) servant d'indices au repérage des structures accentuelles et intonatives qui font l'objet de modélisations phonologiques. La figure 2 illustre concrètement l'évolution des paramètres prosodiques au cours du temps. La figure 3 enfin résume les fonctions susceptibles d'être véhiculées par la prosodie dans les langues à accent.

Prosodie : formes-substance

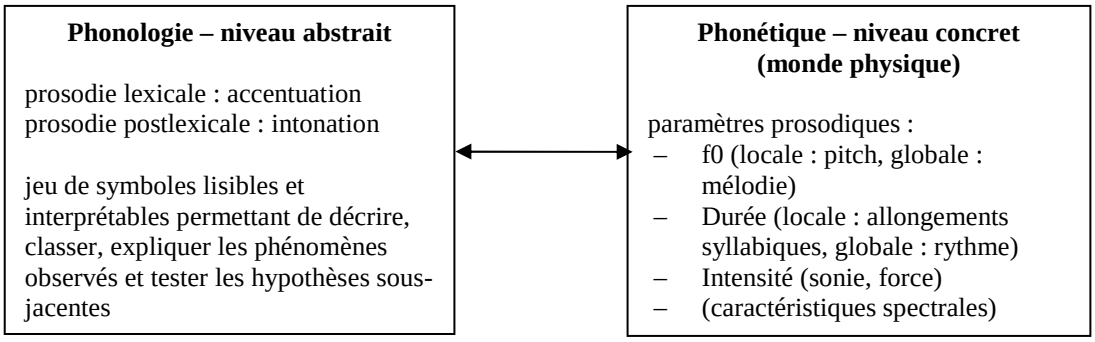


Figure 1 : les domaines phonologique et phonétique de la prosodie

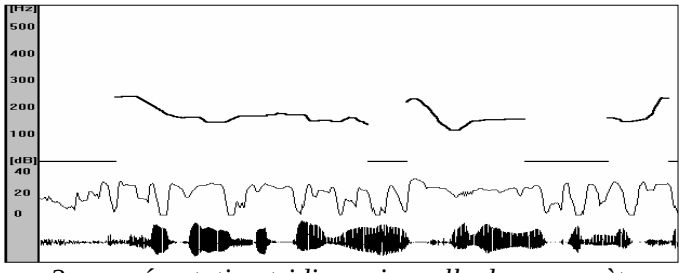


Figure 2 : représentation tridimensionnelle des paramètres prosodiques avec de bas en haut, respectivement, le signal acoustique, l'intensité en décibels, et la fréquence fondamentale en hertz, en repérant par des valeurs nulles les zones non voisées ; sur l'axe des abscisses le temps en millisecondes.

Prosodie : formes-fonctions

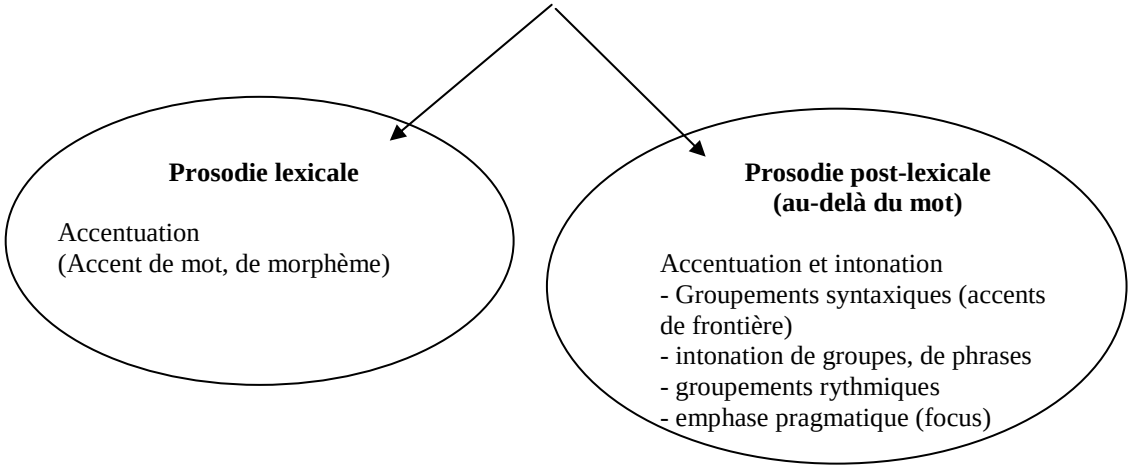


Figure 3. les fonctions lexicales et post-lexicales de la prosodie

Concernant la prosodie du français, la figure 3 appelle une remarque importante : le français faisant partie des langues à accent fixe, la fonction morphologique éventuellement assignée à l'accent reste ténue dans cette langue. L'accent y assumant essentiellement une fonction postlexicale est donc étroitement associé à la structuration intonative qui en

découle directement. Autrement dit, en français, l'accent primaire est un accent de groupe et non de mot, il est localisé obligatoirement sur la syllabe finale du dernier mot d'un groupe syntaxique, éventuellement sur la première syllabe de ce groupe, surtout quand ce dernier est long, il revêt ainsi, au sein du groupe accentuel, une fonction démarcative à droite *(le chat noir, prends-le)* ou à gauche *(Marie et Paul s'amusaient).* Outre l'accent démarcatif, on peut rencontrer des accents dit « secondaires », dont la position est fluctuante à l'intérieur des mots constitutifs d'un groupe accentuel (essentiellement la première syllabe d'un mot quand le contexte phonétique le permet). L'accent secondaire assume une fonction pragmatique (mise en valeur de la partie importante du message : emphase informationnelle ou expressive (*c'est un foulard que je lui ai offert, c'est merveilleux*) et souvent rythmique (équilibrage syllabique entre deux sites accentuels).

3. LE CODAGE PROSODIQUE PROPOSÉ DANS PFC

Une fois résolu, pour la tâche envisagée, le problème de la définition du domaine de la prosodie dans sa dimension phonétique, formelle et fonctionnelle, reste à formuler clairement et explicitement les hypothèses que le codage permettra de tester concernant l'interaction prosodie/schwa, le type de codage le plus optimal pour cette tâche et les outils logiciels associés, enfin la nature des données à coder sur le plan qualitatif et quantitatif. Par ailleurs, un travail de meta-analyse s'impose qui permettra de retenir les aspects les plus opératoires et innovants de la démarche envisagée.

3.1. Hypothèses sur l'interaction schwa/prosodie

Les hypothèses, telles que nous les avons formulées, s'articulent autour de trois thématiques. La première concerne l'interface entre le niveau phonologique profond (*input*) et le traitement phonologique de surface (*output*) ; la seconde s'inscrit dans une problématique rythmique et psycho-acoustique concernant le rôle des constructions prosodiques dans l'intégration perceptive des groupes intonosyntaxiques ; la troisième concerne le conditionnement pragmatique des formes intonatives, *i.e.* l'interaction formes/fonction communicative. Pour ce qui est de la première thématique, la question récurrente posée en phonologie du schwa concerne la façon dont le lien entre forme profonde et forme de surface peut s'établir : la réalisation de surface (dans la parole) correspond-elle dans la structure profonde à un schwa encodé dans le lexique mental, pouvant être élidé en surface ou s'agit-il d'un segment inséré, *i.e.* qui n'a aucune réalité dans la structure profonde ? L'hypothèse que nous défendons est la suivante : si nous sommes face à un processus d'élision, alors la prosodie en porte les traces en surface dans la mesure où la voyelle avant et/ou après le segment élidé est systématiquement différente (durée et/ou patron mélodique) de la même voyelle dans un autre contexte *(ex. c**o**ll'rette vs c**o**lmater*). Au sujet des problèmes d'intégration perceptive, la question est de savoir dans quelle mesure d'une part le nombre de syllabes d'un groupe joue un rôle dans la réalisation d'un « e » muet (longueur du groupe), d'autre part la distance entre le schwa et l'accent primaire influence ou non sa réalisation. L'hypothèse générique que nous posons repose sur le conditionnement rythmique du schwa, elle s'articule autour des deux points suivants : toute chose égale par ailleurs, plus un groupe est long, plus le « e » aura tendance à tomber *(ex. portefeuille vs. porte-monnaie)* ; plus un « e » est proche de l'accent, plus il aura tendance à être élidé, en conséquence la position initiale est plus stable que la position médiane, qu'il s'agisse d'un mot *(**se**couer vs seu**le**ment)* ou d'un groupe *(**je** sais **que** c'est*

comme ça). Concernant enfin le conditionnement pragmatique, deux questions se posent : (i) dans quelle mesure l'organisation topologique d'une séquence (début ou non d'un tour de parole, initiale ou non d'un groupe prosodique) influence la prononciation du « e » ? ;(ii) l'organisation informationnelle joue-t-elle un rôle sur la distribution des « e » réalisés ? À cet égard, les hypothèses que nous défendons se résument comme suit : (i) la position initiale de mot, tout comme le « e » monosyllabique est une position de force de degré variable en fonction de sa distribution topologique dans le discours : le « e » fréquemment élidé au milieu d'un tour étant plus préservé à son initiale *(ça nous **re**garde pas vs. **re**trouver sa joie de vivre)*. (ii) Un « e » est plus facilement effacé à l'initiale d'un segment topical non informationnel que rhématique, porteur d'information, soit les segments suivants pour illustration : *(a) il vient **d'**main avec son frère (b) il vient **de**main avec son frère* en réponse respectivement aux questions : *(a) avec qui vient-il demain ? (b) quand vient-il avec son frère ?*

3.2. Logiciels de codage : quelques remarques

Le problème de fond, sans doute ici plus qu'ailleurs étant donné l'aspect intrinsèquement continu de la matière à décrire, est la quête d'un codage mutualisable, tel qu'il a pu être réalisé au sein de PFC pour le niveau segmental. Avant d'aborder la mutualisation du traitement prosodique en tant que tel (cf. *infra* 3.3), se pose la question de la mutualisation des outils. Nous avons rappelé en 1.1. que nous souhaitions continuer à utiliser PRAAT pour le codage de la prosodie, l'ensemble du codage (schwa, liaison, prosodie, transcription orthographique) étant ainsi disponible dans un et un seul format textgrid sur 4 tires de codage. Or, ce choix ne va pas sans problèmes, le principal étant lié aux fonctionnalités du logiciel : si PRAAT remplit parfaitement sa fonction de logiciel de traitement et d'analyse phonétique, en revanche il n'a pas été conçu pour servir de logiciel de catalogage et de codage, fonction que nous lui faisons portant remplir par l'utilisation de tires parallèles où figurent sur un même plan données et métadonnées (Baude *& al.* 2005). Par ailleurs, le traitement ne peut être que manuel, ce qui suppose un temps de codage significatif et un gros nettoyage des données en aval (contrôle de la transcription) pour lancer les analyses automatiques qui suivront. En revanche, d'autres logiciels proposent déjà des outils d'aide au traitement, c'est le cas de WinPitch (Ph. Martin dév.) qui intègre une fonction d'aide à l'alignement, un ralentisseur de débit pour faciliter le travail de segmentation, la possibilité d'utiliser des balises XML permettant de structurer codage et catalogage, etc. Un transfert vers WinPitch paraîtrait donc souhaitable d'autant que l'interface PRAAT-WinPitch développée depuis peu par Ph Martin permet de récupérer les transcriptions déjà existantes réalisées sous PRAAT. Mais alors d'autres problèmes se posent découlant directement des caractéristiques du logiciel qui n'est toujours pas *open source* aujourd'hui, et qui reste payant pour les collectivités, ces deux caractéristiques étant totalement en contradiction avec une philosophie de mutualisation. Bref, la question suivante, récurrente en TAL et en linguistique de corpus, reste ouverte : comment trouver le meilleur compromis entre des objectifs de partage et de mutualisation d'un côté, un traitement plus pointu, intégrant des outils d'aide à la transcription de l'autre ?

3.3. Méthode et protocole de codage

Concernant la démarche méthodologique d'ensemble, si l'opposition entre méthode hypothético-déductive et stratégie inductive est souvent posée comme allant de soi, il s'agit

dans la pratique d'un artefact. En effet, de même que l'éclairage des hypothèses par les données suppose un va-et-vient étroit entre hypothèses et données, une méthode uniquement inductive est illusoire : on ne cherche pas n'importe quoi, n'importe où, des hypothèses posées préalablement orientent drastiquement la recherche. Soit le cheminement suivant (figure 4) :

- Dans une phase initiale, un premier jeu d'hypothèses va permettre de mettre en place un protocole de codage préalablement à la sélection des contextes à coder et à leur codage effectif. Une première exploration des données va nous amener à préciser nos hypothèses et à modifier notre protocole le cas échéant, notamment à restreindre ou étendre les champs de codage et les informations qui s'y déploient.
- Ce n'est que lorsque cette phase initiale est vraiment stabilisée que des requêtes automatiques peuvent être lancées afin de fournir des résultats statistiques qualitativement et quantitativement représentatifs pour l'analyse des données, conformément aux hypothèses stabilisées.

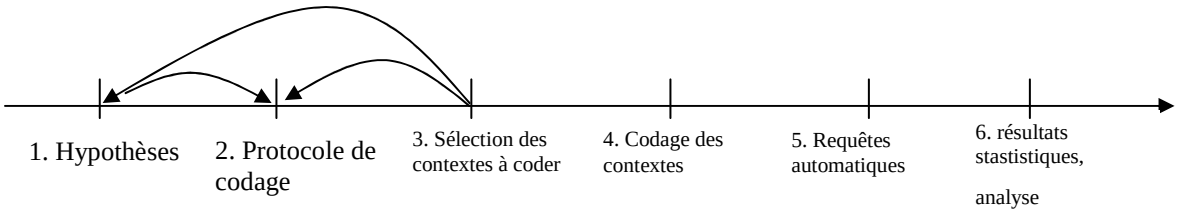


Figure 4 : la démarche sur corpus, des hypothèses aux données, des données aux hypothèses

Nous ne reviendrons pas sur les hypothèses telles qu'elles ont été posées en 2.1, attardons-nous en revanche sur la stratégie de codage retenue : type et nature du codage, sélection des contextes à coder. Nous avons opté pour un codage générique quel que soit le type de production transcrite (lecture de texte, conversation dirigée, conversation libre) en sachant que certains marqueurs ne seront jamais utilisés dans des types de production et abondants dans d'autres (voir par exemple les marqueurs codant les pauses d'hésitations inutiles pour la transcription du mode <lecture de texte>).

Le codage se subdivise en deux ensembles : un codage standard réalisable par des non spécialistes entraînés (2 heures d'apprentissage environ) et un codage étendu qui nécessite en partie une expertise phonétique et des logiciels dédiés. En pratique, l'architecture globale du protocole comprend 9 champs, les deux premiers fixant nos domaines de codage prosodique (la syllabe et le mot), les 7 autres fournissant les informations prosodiques exploitées pour les tests de nos hypothèses. Précisément, les 6 champs du codage standard se décomposent comme suit : le **champ 1** indique le nombre de syllabes prononcées du mot en cours de traitement, le **champ 2** code le rang de la syllabe analysée dans le mot, le **champ 3** renseigne sur la perception de la syllabe : proéminente ou non, Le **champ 4** indique des variations de durée syllabique éventuelles, Le **champ 5** donne des informations sur les occurrences des pauses produites et leur nature. Précisons que, pour l'heure, ce champ reste relativement sommaire par rapport aux descriptions que l'on a pu proposer sur les types de dysfluences rencontrées dans le discours (pour une étude récente, voir Assié, 2005). Nous distinguons les pauses silencieuses courtes et longues (seuil de référence = 300 ms), les « euh » d'hésitation et les répétitions syllabiques (*c'est*

in/incroyable). Le **champ 6** permet d'élargir le domaine de codage au sein du discours (syllabe initiale ou non de tour de parole). Le protocole étendu pour sa part renseigne sur les corrélats psycho-acoustiques des objets phonologiques dans les distributions observées. En pratique, il s'agit de tenir compte du niveau fréquentiel des noyaux syllabiques (champ 7), de la configuration mélodique des contours syllabiques, pente montante ou descendante, forte ou faible (champ 8) et de l'intensité associée (champ 9).

Concernant enfin la sélection des contextes, rappelons que le codage de la prosodie se distingue significativement du codage segmental dans la mesure où les unités à coder ne sont plus des lexèmes/graphèmes mais inévitablement des syllabes. La segmentation syllabique est en effet incontournable si on veut disposer d'informations prosodiques minimales certes mais suffisamment exhaustives pour pouvoir en tirer des informations exploitables. Un tel travail est une tâche lourde et fastidieuse et, de fait, non dénuée de risques d'erreurs qu'il faut pouvoir, à terme, contrôler au maximum (cf. *infra*, conclusion). En conséquence, si plusieurs stratégies de sélection des passages sont *a priori* envisageables, peu s'avèrent tenir la route bien longtemps. En particulier, on pourrait souhaiter maintenir le parallélisme avec le codage schwa et le codage liaison en proposant la transcription aléatoire d'une portion du texte et d'une minute de chaque entretien par exemple. Cette procédure peut fournir des résultats fructueux à condition que les occurrences recherchées soient suffisamment fréquentes ce qui est souvent loin d'être le cas. Bref, il est vite apparu qu'une démarche non contrôlée ne donnerait pas de résultats pertinents, étant donné notre objectif particulier qui vise à mettre en relation la présence/absence du schwa avec la prosodie dans des contextes bien circonscrits. Une deuxième solution alors consiste à sélectionner attentivement les passages à coder en ayant constamment à l'esprit les types de requêtes que nous souhaitons effectuer par la suite, de manière à tester des hypothèses phonologiques clairement formulées au préalable. Bien entendu, nous avons opté pour cette seconde approche sans pour autant verser dans un optimisme débordant. La parole spontanée reste par définition spontanée sans garantie aucune de fournir tous les types de contextes que nous nous proposons d'envisager, d'où la nécessité d'expériences complémentaires en laboratoire (cf. *infra*, conclusion). Environ cinq locuteurs sont retenus pour chaque point d'enquête soumis au codage prosodique et la sélection s'établit sur la base de la qualité du signal, la présence de variations mélodiques bien nettes, enfin la richesse potentielle des contextes. Les passages à coder du texte sont préétablis et favorisent l'étude des sites d'insertion attendue de schwa et les quelques instances de monosyllabes où l'absence de schwa est envisageable. Pour les conversations (dirigée et libre), nous préconisons le codage d'une minute de chaque type conversationnel pour chacun des témoins retenus. Pour finir, précisons que les passages sélectionnés sont discontinus et sont choisis essentiellement de façon à mettre en valeur les patrons prosodiques qui interpellent le codeur, *i.e.* des patrons que l'on qualifiera de « marqués » par rapport à l'intuition du codeur de ce que peut être une prononciation non marquée ou standard.

4. CONCLUSION

Dans cette article, nous avons souhaité baliser les grands axes de ce que peut être une base de données phonologique dans sa composante prosodique en interaction avec la couche segmentale. La démarche unifiante de ce travail est la suivante : la prosodie permettrait de confirmer certaines hypothèses au détriment d'autres concernant la prononciation variable

des segments, en l'occurrence le schwa (ou « e » caduc ou encore « e » muet). Elle constituerait notamment un indice précieux pour confirmer l'hypothèse d'un processus d'élision dans la parole d'un segment encodé dans le lexique mental des locuteurs ; à l'inverse, la prosodie peut agir comme un déclencheur pour la réalisation du schwa, il existe en particulier une relation entre les structures rythmiques et la prononciation du « e » caduc. Evidemment, pas de théorie sans méthodes et sans outils surtout quand il s'agit d'explorer des données à grande échelle, c'est bien là le cœur du propos. Cet aspect fondamental nous a amenées à exposer notre programme méthodologique autour de 3 points : (1) la mise en place d'un premier jeu d'hypothèses pour le développement d'une série de requêtes en cours d'implémentation, (2) notre protocole de codage tel qu'il est stabilisé aujourd'hui après deux ans de réflexions et d'essais sur corpus, (3) la stratégie de sélection des contextes pour l'analyse. Si nous attendons beaucoup de cette première phase qui nous permettra de préciser les processus de variation dans lesquels interagissent effectivement les niveaux segmental et suprasegmental, et si un effort colossal a été effectué pour une sélection rigoureuse des contextes sur différents points d'enquête, encore une fois, nous restons réalistes quant à leur représentativité tant d'un point de vue quantitatif que qualitatif. En d'autres termes, il est évident que pour l'ensemble des contextes dont nous souhaitons rendre compte, il y aura nécessairement des trous, les données étant parfois étrangement silencieuses... Deuxième problème : celui de l'outillage. La tâche de transcription manuelle est un travail long et fastidieux, nous l'avons dit, les erreurs inévitables. Des outils d'assistance au codage (syllabation semi-automatique, détection de proéminences) et des outils de contrôle sont donc indispensables. Dans cette optique, le développement d'un contrôleur pour la vérification des étiquettes est en cours de discussion avec l'équipe PFC du laboratoire Modyco (coll : A. Tchobanov, R. Walter). Troisième restriction : certaines mesures fines comme celle de l'allongement vocalique éventuel suite à la chute du schwa demandent un contrôle drastique des variables manipulées, d'où la nécessité de passer le relais à ce stade à la phonologie de laboratoire. D'une manière générale, c'est sans doute cette collaboration soutenue entre approche de terrain et approche de laboratoire qui permettra de reconfigurer le paysage de la linguistique contemporaine en proposant des outils pour tester les hypothèses des linguistes et alimenter ainsi les vieux débats, mais aussi pour en ouvrir de nouveaux.

5. REFERENCES

Assié D. 2005. *Analyse syntaxique automatique de corpus oraux retranscrits.* Mémoire de DEA en Sciences du langage, Toulouse le Mirail.

Durand J., Lyche Ch. 2003. ''Le projet 'Phonologie du Français Contemporain' (PFC) et sa méthodologie'' in Delais-Roussarie E. et Durand J. (eds.). *Corpus et variation en phonologie du français*. Toulouse : PUM, 213-276.

Baude O. M. Jacobson, Tchobanov A. Walter R. 2005. 'Interopérabilité des corpus sonores : le cas des corpus en français', *Phonological Variation : the Case of French,* University of Tromsø, Norway, 25-27 August 2005, http://www.projet-pfc.net/.

Lacheret A., Beaugendre F. 1999. *La prosodie du français,* Paris, Editions du CNRS.

LE WEB ENTRE CORPUS LINGUISTIQUE ET ABERRATION LINGUISTIQUE : LE CAS DES PRONOMS ANGLAIS DE PREMIÈRE PERSONNE

Antoine Consigny
EA1339, Université Marc Bloch - Strasbourg 2, anconsig@umb.u-strasbg.fr

1. INTRODUCTION

Le but de cet article est de présenter une étude des pronoms personnels et du pronom réfléchi de première personne en anglais, grâce à concordances faites à partir du web. L'approche utilisée est la grammaire systémique fonctionnelle de Halliday. Les différentes utilisations sont analysées dans ce cadre, et il apparaît quelques anomalies dans l'utilisation des pronoms anglais de première personne sur Internet. On en vient alors à se demander si le web peut être considéré comme corpus linguistique ou s'il est plus proche de l'aberration. Ces deux visions opposées ont été exprimées plus ou moins directement au cours de ces journées, en particulier par Véronis (voir ce volume) et par Williams respectivement.

2. LES DONNÉES

Deux concordances ont été effectuées grâce au logiciel Webcorp (Morley et *al*, 2003 ; Renouf, 2004). Ce logiciel permet, en utilisant des moteurs de recherche du web, de réaliser des concordances de mots ou de structures en utilisant le web comme corpus. Il est possible de restreindre à un domaine particulier voire à des types de journaux (« UK Broadsheets » par exemple cherchera dans les textes en ligne de grands journaux britanniques). Pour cette étude, au contraire, j'ai préféré ne donner aucune restriction de ce type, ce qui fait que les sites recherchés pouvaient venir de tout le web[1].

La quantité de contexte autour du mot clé peut également être modifiée ; dans l'étude, la phrase a été choisie comme étant *a priori* suffisante, et surtout car elle est aussi l'unité linguistique choisie pour l'analyse des données.

La première concordance a pour mot clé *myself*, et la deuxième, *me*. Pour le pronom *I*, les concordances précédentes ont été utilisées, ce qui donne suffisamment de lignes. Les données sont les suivantes :

	I	*me*	*myself*
Nombre de lignes	256	235	374[2]

Tableau 1 : I, me et myself, données brutes

Dans la suite de l'article, les exemples ont été laissés tels quels, sans corriger les éventuelles fautes d'orthographe ou de grammaire ni enlever les italiques où il y en avait. Pour plus de lisibilité, des italiques ont toutefois été ajoutés pour le(s) mot(s) analysé(s).

3. ANALYSE SYSTEMIQUE FONCTIONNELLE

3.1. Bref rappel de l'analyse systémique fonctionnelle : les fonctions du langage

Pour Halliday, une proposition est utilisée pour remplir toutes les fonctions du langage ; à chacune correspond un type de sens préférentiel.

Pour la fonction textuelle, il considère la proposition en tant que message, c'est-à-dire en tant que quantité d'information dont l'organisation est importante. La structure principale pour ce type de sens est la structure thématique en Thème et Rhème. Le Thème est la première partie de la proposition, le point de départ du message (voir aussi Francis, 1989).

Pour la fonction interpersonnelle, la proposition est vue comme un échange, entre les participants. La structure utilisée ici est Mode (composé de Sujet et Elément Conjugué) - Résidu, c'est-à-dire ce qui ne change pas avec la polarité (positive ou négative) ou avec la forme (affirmative, interrogative etc.). Le Sujet (au sens de sujet de l'élément conjugué) est l'élément désigné par le locuteur comme responsable de la validation de l'énoncé.

Pour la fonction idéationnelle, la proposition est vue comme une représentation, c'est-à-dire la construction d'un procès dans l'expérience humaine. Dans ce sens, on parle d'Acteur, de Procès et (éventuellement) d'autres parties, comme les objectifs, attributs (dans le sens de ce qui caractérise le Sujet) etc. L'Acteur est l'élément actif dans le procès, celui qui fait l'action.

L'avantage de ce type de distinction est de rendre compte des différentes facettes que contient souvent la définition du sujet : « ce qui fait l'action » ; « ce dont on parle » ; « ce qui est repris dans une reprise par auxiliaire », qui ne correspondent pas toujours à la même chose. Elle sera particulièrement utile pour les pronoms étudiés, surtout pour le couple *I-me*

3.2 Exemple : étude d'une phrase sur trois niveaux

On prendra deux phrases données en exemple pour la division en Thème + Rhème par Halliday (1994, p. 34), et en ajoutant les autres divisions :

I	*caught*		*the first ball*
Thème	Rhème		
Mode		Résidu	
Sujet	Elément conjugué	Prédicat	Complément
Acteur	Procès		Objectif

I	*was*	*beaten*	*by the second*
Thème	Rhème		
Mode		Résidu	
Sujet	Elément conjugué	Prédicat	Adverbial
Objectif	Procès		Acteur

Figure 1 : étude de deux phrases simples

Pour Halliday, dans une phrase déclarative non marquée, le plus souvent le Thème est aussi Sujet de l'élément conjugué et Acteur du procès (matériel). Il existe ensuite des déviations par rapport à ce schéma de base, suivant le type de phrase, etc.

Les divisions faites ci-dessus permettent de montrer que ce n'est pas forcément le Sujet (grammatical) qui fait l'action. De même il serait possible de montrer que le Sujet n'est pas forcément le Thème. Dans les phrases ci-dessus, un locuteur donne une information sur lui-même, d'où le pronom *I* en position thématique et sujet. On reprend cette analyse dans la partie suivante avec les pronoms anglais *I*, *me* et *myself*.

4. LES PRONOMS ANGLAIS DE 1ÈRE PERSONNE : LES UTILISATIONS « NORMALES »

4.1. Le pronom Sujet *I*

Celui-ci ne pose pas de problème, il ne se trouve dans les concordances qu'avec la fonction de sujet, et le plus souvent en position thématique même si ce n'est pas toujours le cas. Sur le plan interpersonnel, il fait référence à l'énonciateur lorsque celui-ce fait référence à lui-même (exemple 1) :

Ex. 1 : *I* took the manuscript with me on a trip to the Dolores River

I	took		the manuscript	with me	on a trip to the Dolores River
Thème	Rhème				
Sujet	Elt conjugué	Prédicat	Complément	Adverbial	Adverbial
Acteur	Procès matériel		Objectif	Circonstance	Circonstance

Figure 2 : Exemple 1

Ex. 2 : If you post one in the comments, (…) *I*'ll send a treat as a thank you!

If you post one in the comments,	I	'll	send	a treat	as a thank you
Thème	Rhème				
Adverbial	Sujet	Elt conjugué	Prédicat	Complément	Adverbial
Circonstance	Acteur	Procès matériel		Objectif	Circonstance

Figure 3 : Exemple 2

On voit que le pronom ne change pas suivant sa place : il peut être thématique ou rhématique, Acteur ou non (*cf.* le deuxième exemple de la section 3.2.), même si c'est plus souvent le cas, mais il est toujours Sujet de l'élément conjugué, donc dans le Mode.

4.2. Les cas obliques : *me* et *myself*

Pour les fonctions autres que Sujet, les deux autres pronoms peuvent se trouver. Il s'agit donc de voir ce qui les différencie, et quand l'un est privilégié par rapport à l'autre.

4.2.1 ME Les exemples 3 à 6 montrent les emplois normaux de *me*.

Ex. 3 : He had been expecting *me*.

Ex. 4 : Everything on this site is created by *me*…

Ex. 5 : I can't speak for anyone else, but for *me* novel writing is very incremental, like …

Ex. 6 : … here comes Abbey in the ranger power-boat to watch *me* run Badger Creek Rapid.

On trouve le pronom essentiellement en position rhématique (exemples 3 et 4), quoique quelques lignes le montrent à l'inverse, comme dans l'exemple 5. C'est cependant rare (1,7 % contre 98,3 % en position rhématique). De même il est le plus souvent Complément (d'objet) du Prédicat (68,5 % des cas), ce qui fait de l'exemple 3 le cas typique. Le reste du temps, comme dans les exemples 4 et 5, il est essentiellement (27,2 % du total) Adverbial (Adjunct). Il fait donc typiquement partie du Résidu.

Dans l'exemple 5, on considère la proposition « but… incremental ». Le pronom se trouve dans un Thème complexe composé d'une partie textuelle (*but*, qui fait le lien avec la proposition précédente et assure la cohésion de la phrase) et d'une partie topicale[3] (expérientielle/idéationnelle) ET interpersonnelle, *for me*. En effet, ce syntagme prépositionnel est tout à la fois « ce dont on parle » et un des participants dans le discours.

Enfin, l'exemple 6 montre que *me* peut être Acteur d'un procès matériel. C'est cependant un cas un peu particulier, quoique assez fréquent. Il est Objectif dans la proposition principale dont le procès est watch, mais Acteur dans la subordonnée, non conjuguée (et donc sans Sujet) run Badger Creek Rapid. Ce cas où *me* est Acteur est bien plus fréquent que les propositions à verbe au passif (exemple 4) ; il ne remet cependant nullement en cause le statut du pronom, et est la seule possibilité, surtout avec un verbe non conjugué. Il en serait différemment si on avait une subordonnée conjuguée (du type subordonnée en that, le Sujet serait alors de nouveau I).

Dans tous ces cas, le pronom me sera utilisé lorsque le Sujet n'est pas l'énonciateur, c'est à dire qu'il n'est pas *I*[4].

4.2.2 MYSELF Les exemples suivants donnent une idée du comportement « normal » trouvé dans la concordance :

Ex. 7 : It made me despise *myself* and got me to thinking …

Ex. 8 : sometimes I think I am a king in spite of *myself*.

Ex. 9 : I am not particularly fond of goat cheese *myself*.

Ex. 10 : I *myself* become the wounded person

Ex. 11 : … I had to learn and relearn-basic life skills, learning how to be *myself*.

Comme le précédent, ce pronom a dans la concordance une préférence pour la position rhématique, même si elle est moins marquée (19,5 % contre 80,5 %). Une des raisons pour la plus grande fréquence des emplois thématiques est peut-être la répétition du sujet sous forme réfléchie, comme dans l'exemple 10. L'autre raison, plus flagrante, est le grand nombre d'emplois où *myself* est employé « contre nature » dans une position thématique (voir section 5.2.), ce qui a changé les résultats.

Ses fonctions principales en position thématique sont les mêmes que pour *me*, les fonctions (cas obliques) d'Objectif (exemple 7), de Circonstance (exemple 8), Identifieur/Attribut (exemple 11) etc.

Une autre utilisation de *myself* habituelle, et tout à fait fréquente est l'emphase, où il est utilisé en reprise de l'Acteur/Senteur/Porteur (exemples 9 et 10). Dans ces cas, le pronom réfléchi est considéré comme Acteur (ou équivalent) répété.

On voit que le pronom réfléchi s'il apparaît dans une proposition conjuguée apparaîtra après le Sujet *I*, et si la proposition est non conjuguée (exemple 7), le pronom *me* est présent ou il doit y avoir une référence explicite à l'énonciateur dans la proposition ou tout au moins dans la phrase (exemple 11).

Un cas diffère légèrement de ce schéma, c'est celui de l' « hypercorrection » (Wales, 1996). C'est lorsque l'énonciateur utilise *myself* alors qu'il n'y a pas eu de référence

suffisante à lui-même dans le contexte précédent. Dans ce cas, on pourrait utiliser *me* à la place. L'exemple suivant le montre :

> Ex. 12 : And you shouldn't say "to my wife and *myself*". The only correct way so say this is, "The IRS sent the refund check to my wife and me".

Cet exemple métalinguistique montre à la fois que *me* est possible et que *myself* se trouve. Le reste du site commente d'ailleurs cette utilisation, qui est souvent refusée (voir encore Wales, 1996).

4.3. Conclusion provisoire : les différences et affinités entre les trois pronoms

Le tableau comparatif suivant donne une indication sur la répartition des trois pronoms dans les concordances, suivant les différentes réalisations possibles des trois fonctions linguistiques de la grammaire fonctionnelle.

Fonction		*I*	*me*	*myself*
Textuelle	Thème	207 (80,9 %)	4 (1,7 %)	73 (19,5 %)
	Rhème	49 (19,1 %)	231 (98,3 %)	301 (80,5 %)
Interpersonnelle	Sujet	256 (100 %)	5[5] (2,1 %)	68[6] (18,2 %)
	Complément	0	161 (68,5 %)	132 (35,3 %)
	Adverbial	0	64 (27,2 %)	168 (44,9 %)
	Attribut	0	5 (2,1 %)	6 (1,6 %)
Idéationnelle	Acteur/Porteur/Senteur	253 (98,8 %)	44 (18,7 %)	112 (29,9 %)
	Objectif	1 (0,4 %)	63 (26,8 %)	117 (31,3 %)
	Circonstance	0	48 (20,4 %)	87 (23,3 %)
	Bénéficiaire/Cible	2 (0,8 %)	75 (31,9 %)	52 (13,9 %)
	Identifieur	0	5 (2,1 %)	6 (1,6 %)
Total lignes		256	235	374

Tableau 2 : répartition de I, me et myself par fonction

Tout d'abord, il y a une grosse différence en termes de Thème-Rhème, le premier étant le plus souvent réalisé par *I* et le deuxième par les deux autres pronoms.

Dans la répartition en Mode et Résidu, la différence se retrouve, de façon encore plus nette : *I* ne se trouve que dans le Mode, les deux autres très majoritairement dans le Résidu.

La référence est la même pour les trois : il s'agit de l'énonciateur à chaque fois. Ils ont donc également la même fonction interpersonnelle de renvoi à l'énonciateur puisqu'il fait partie, par définition, des acteurs de l'interaction langagière.

Cependant, cette référence ne se réalise pas tout à fait dans les mêmes conditions pour les deux pronoms en cas obliques : *me* sera utilisé lorsque le sujet de la proposition a un autre référent que l'énonciateur, alors que *myself*, en tant que pronom dit « réfléchi », se trouvera naturellement dans le cas contraire (voir les exemples précédents). On trouve donc un lien privilégié *I-myself* pour ce qui est de la cohésion à l'intérieur d'une proposition ; plus rarement *I-me* ou *me-myself*. Avec *me*, c'est plutôt une relation du type X-*me*, où X est différent de *I*.

En termes de participants au Procès, *I* en est un, et le plus souvent un participant actif (Acteur, Porteur etc.). *Myself* peut n'être que la répétition de *I*, c'est-à-dire la répétition d'un participant déjà mentionné dans la proposition, alors que *me* sera généralement un des participants par opposition à d'autres.

5. LES UTILISATIONS « ANORMALES »

Par « anormal » je ne veux surtout pas énoncer un jugement de valeur. Il n'est pas entendu par là que les emplois trouvés ne sont pas acceptables, mais qu'ils ne correspondent pas à ce qui est attendu dans le langage dit « correct » ou soigné, ni aux soi-disant incorrections dues au langage familier ou plus relâché.

5.1. *Me* sujet

Il y a 5 exemples où *me* se trouve Sujet de l'élément conjugué. Deux cas distincts se trouvent, comme le montrent les deux exemples :

> Ex. 13 : *me* an Daniel (who sux at html) are gonna fix them up reel nice like.
>
> Ex. 14 : But the me who sent his first story into Asimov's seventeen years ago couldn't answer the question, "Write a story for Asimov's" the same way the me of today could.

Dans le premier cas, le pronom fait partie d'un Sujet composé où il est coordonné avec un nom ; dans le deuxième cas, *me* est utilisé, quoi qu'il fasse bien référence à l'énonciateur, comme substantif (d'où l'article défini the). Le premier, s'il est considéré comme incorrect, n'en existe pas moins en anglais « relâché » et peut donc tout à fait entrer dans une description de l'anglais. On devrait en effet avoir Daniel and *I* mais l'usage ici (c'est confirmé dans le reste du site) diffère quelque peu du « bon anglais ». Le deuxième n'entrera pas dans une description des pronoms de l'anglais, sauf pour dire que l'anglais a une certaine souplesse et qu'un pronom peut parfois changer de catégorie, avec un changement de sens, de même que moi en français peut devenir substantif (le moi d'aujourd'hui…). Compte tenu de la remarque faite en début de section, ces deux emplois ne sont pas « anormaux ».

5.2. *Myself*, un texte de blog

Il y a quelques cas où les deux pronoms cas obliques font office de sujet, mais il s'agit alors de propositions dites « non finies » (verbes non conjugués), comme pour *me* dans l'exemple 6 ; ils ne sont pas problématiques. À l'inverse, dans cette section, on étudie l'utilisation de *myself* essentiellement mais pas uniquement en fonction de Sujet de l'élément conjugué.

L'utilisation se trouve dans deux sites, tous deux des blogs. Dans le premier, il n'y a qu'un cas :

> Ex. 15 : *Myself* always thinks with ego about what if it was "the one", …

Sur cet exemple, on peut se demander s'il s'agit d'une erreur ou d'un choix délibéré (Consigny, 2005). Il n'en demeure pas moins qu'il apparaît dans la concordance. En termes hallidayens, *myself* est ici Thème, Sujet et Senteur (d'un procès mental), ce qui ne correspond en rien aux emplois « normaux ».

Dans le deuxième blog l'auteur se fait appeler *Myself*. Le problème que cela pose est que l'usage de *myself* et celui des pronoms personnels et des possessifs n'est pas cohérent. Lorsqu'on regarde le site complet, on se rend compte que l'appellation par l'auteur de *myself* ne fait aucun doute. On trouve même l'adjectif possessif de première personne :

> Ex. 16 : So today way *Myself*'s second to last day at work.
>
> Ex. 17 : So being that *Myself* has a personal interest in all of this, here's *my* 2¢ worth.

Ex. 18 : said friend has been a real help in getting D & *Myself* set up here, took *us* to brunch.

Il y a d'autres cas cependant où l'auteur utilise *I* à la place de *myself*. Les deux se trouvent parfois côte à côte dans la même phrase, dans un ordre ou l'autre :

Ex. 19 : *Myself* says, *I* just want to go about with *my* daily stuff, …

Ex. 20 : And *I* further don't care of any church recognizes it, *Myself* is more concerned what the (…) benefits are.

On se trouve donc devant un problème, car le pronom *I* se voit utilisé de façon anaphorique, en reprise du nom *Myself*. Cet emploi est déjà en soi inhabituel, puisqu'un énonciateur en principe parle de lui à la première personne (mais voir Consigny, 2005), et dans ce blog il alterne première et troisième. Il va jusqu'à utiliser le pronom réellement anaphorique de troisième personne *he*, et le possessif correspondant :

Ex. 21 : … there are somethings about *Myself* that *he* doesn't *want* to let anyone know about.

Ex. 22 : *Myself* then has 15 dans to pack up *his* life and move to…

En ce qui concerne plus précisément les fonctions de *myself*, on peut les résumer dans le tableau suivant (en comptant l'unique exemple du premier blog) :

Fonction		myself
Textuelle	Thème	64 (92,8 %)
	Rhème	5 (7,2 %)
Interpersonnelle	Sujet	63 (91,3 %)
	Complément	3 (4,3 %)
	Adverbial	3 (4,3 %)
	Attribut	0
Idéationnelle	Acteur/Porteur/Senteur	61 (88,4 %)
	Objectif	1 (1,4 %)
	Circonstance	4 (5,8 %)
	Bénéficiaire/Cible	3 (4,3 %)
	Identifieur	0
Total lignes		69

Tableau 3 : répartition des emplois de myself par fonction

5.3. Discussion

5.3.1 SUR LES EMPLOIS Tout d'abord, on peut dire que les utilisations « anormales » de *me* sont finalement une question de registre plutôt que de langue. En d'autres termes, ces utilisations font partie de la langue, même si elles sont considérées comme « incorrectes » par certains dans le cas de sujet composé. Elles sont certainement plus acceptables à l'oral qu'à l'écrit. Pour ce qui est de *me* utilisé comme nom, il ne pose pas vraiment problème.

En ce qui concerne *myself*, la réponse est moins aisée. Le mot n'est pas utilisé comme pronom dans le texte de blog, sauf dans un ou deux cas, et il aurait pu être purement et simplement supprimé de la concordance. Cependant, la concordance a été étudiée sans parti pris de « correction », c'est-à-dire que toutes les occurrences qui avaient un sens ont été étudiées. Ce que le blog montre, par contre, c'est que l'auteur utilise le langage à sa propre convenance, sans se soucier du reste. C'est la raison pour laquelle on trouve *I* dans

une fonction anaphorique de répétition lexicale, c'est-à-dire avec une fonction textuelle, de cohésion (Halliday et Hasan, 1976).

Sur un plan interpersonnel comme sur le plan purement syntaxique, il y a confusion chez l'énonciateur entre un lui détaché, à la troisième personne y compris dans la reprise par pronoms, et un lui qui ne l'est pas, ce qui le met donc en concurrence avec les possessifs et pronoms de première personne (y compris pluriel lorsqu'il parle de lui et d'autres).

5.3.2 SUR LA CONCORDANCE On peut revenir sur les statistiques de la section 4.3., où les usages étaient répartis par fonction. Pour ce qui est de *me* et de *I*, elles seront inchangées, puisque les utilisations des deux pronoms personnels sont en fin de compte acceptables dans leur contexte, et font partie de la langue. Par contre, on obtient des statistiques plus fiables sur les emplois attendus de *myself*, puisqu'on supprime ceux qui ne le sont vraiment pas.

Fonction		*myself*
Textuelle	Thème	9 (3,0%)
	Rhème	296 (97,%)
Interpersonnelle	Sujet	5 (1,6%)
	Complément	129 (42,3%)
	Adverbial	165 (54,1%)
	Attribut	6 (2,0%)
Idéationnelle	Acteur/Porteur/Senteur	51 (16,7%)
	Objectif	116 (38,0%)
	Circonstance	83 (27,2%)
	Bénéficiaire/Cible	49 (16,1%)
	Identifieur	6 (2,0%)
Total lignes		305

Tableau 4 : répartition des fonctions de myself hors texte de blog

On trouve alors des statistiques relativement proches entre les emplois de *me* et de *myself*, ce qui est cohérent puisque la différence principale entre les deux pronoms est dans la présence ou non de la coréférence (un renvoi explicite à l'énonciateur pour *myself*).

5.3.3 SUR LE WEB COMME CORPUS Il reste cependant un certain nombre d'autres problèmes sur ces statistiques, qui sont dus au type de corpus. En particulier, il y a des sites dans lesquels les mêmes emplois sont répétés. Par exemples, un site qui donne des paroles de chansons peut avoir plusieurs lignes avec le même emploi, soit parce qu'une même phrase est répétée (un refrain) ou reprise dans le titre et dans le texte. Dans le même ordre d'idée, il peut y avoir les mêmes textes sur plusieurs sites, par exemples des textes littéraires ou poétiques.

Webcorp propose une solution possible, avec l'option de ne sélectionner qu'une ligne par site, pour ne pas avoir ce problème. Cependant, cela occulte aussi les sites qui offrent une variété d'emplois, et on risque ainsi de passer à côté de beaucoup.

Un dernier problème est posé par les résultats bruts de la concordance. Puisque le web est considéré dans son entier, il peut arriver (ce qui est le cas ici) que des emplois se glissent dans la concordance alors qu'ils ne sont pas corrects ou ne correspondent pas à ce qu'on attend. On revient alors aux critiques formulées par divers linguistes sur l'utilisation des corpus, en les appliquant au web. Cruse en 1986 dit de la linguistique de corpus que « le linguiste n'a aucun contrôle sur ce qu'il étudie ». Owen (1996) se méfie de certaines utilisations pédagogiques, en particulier sur l'utilisation des corpus par des apprenants,

s'ils n'ont pas de connaissances suffisantes pour aborder une concordance. Sur cette utilisation, on peut d'ailleurs consulter Watson-Todd (2001).

D'un autre côté, on peut se demander si les emplois qui ne sont pas acceptables pour l'instant ne le deviendront pas, et le web est alors une bonne source de productivité linguistique. On pourra rappeler en particulier les caractéristiques orales de la langue utilisée dans de nombreux sites personnels. Pourtant l'auteur du blog peut bénéficier de certains avantages de l'écrit, en particulier la possibilité de se corriger, le temps de choisir son vocabulaire etc. Cela fait du blog un genre hybride et potentiellement lieu de créativité langagière importante. Celle-ci fait d'ailleurs l'objet d'études, par exemple Smyk (2005) qui regarde les innovations lexicales dans les blogs. Sur les blogs comme genre, on pourra aussi voir Herring et al (2004) ; Marlow (2004) ; Squire (2003).

6. CONCLUSION

Cette communication avait pour but d'étudier les différences d'emploi des trois pronoms personnels de première personne *I*, *me*, *myself*. On a pu ainsi montrer, en utilisant la grammaire fonctionnelle développée par Halliday, qu'ils remplissaient des fonctions différentes dans les propositions, même s'ils ont la même référence quant au participant à l'interaction langagière, l'énonciateur.

Elle a permis également de mettre en lumière un certain nombre d'usages inattendus, et *a priori* inacceptables. En effet, *myself* est utilisé par certains internautes comme autre chose qu'un pronom qui sert à confirmer que l'énonciateur parle de lui. De plus, il occupe parfois la fonction de Sujet, ce qui entre les trois pronoms est normalement réservé à *I*.

On a ainsi pu relever des problèmes liés à la méthodologie, et aussi au corpus lui-même. Webcorp permet d'obtenir des concordances à partir du web, mais on a peu de contrôle sur les résultats, d'où la question : le web est-il un bon corpus, et une bonne source de données linguistiques ?

Deux réponses partielles peuvent être données, selon le type de résultat que l'on recherche. Tout d'abord, il est possible, si elle n'est pas trop longue, d'éditer manuellement la concordance, en enlevant les lignes « indésirables », que ce soient les emplois idiosyncrasiques ou les emplois répétitifs (voir Lüdeling et *al*, à paraître). Sinon, on peut également essayer de faire des sub-corpus par genres (voir *idem*, et Sharoff en soumission). Il devrait alors devenir possible de voir si les emplois inattendus qui proviennent de blogs sont plus répandus dans ceux-ci, ou si les blogs sont effectivement des lieux de créativité linguistique.

7. NOTES

1. Ceci a pour inconvénient d'inclure dans la recherche des sites indésirables, notamment des sites pornographiques ou assimilables. Les lignes provenant de ces sites ont été retirées des décomptes.
2. Le nombre plus grand s'explique peut être par le fait que la concordance de *myself* a été réalisée plus tôt que l'autre, à un moment où le logiciel utilisait simultanément plusieurs moteurs de recherche, ce qui n'est plus le cas.
3. C'est à dire le sujet de l'énoncé, ce dont on parle ; la traduction est de D. Banks.
4. Quelques exceptions existent, notamment avec la préposition *with*, voir exemple 1. Voir aussi Quirk et al. (1985) pour une description plus détaillée des emplois et exceptions.
5. Dont 3 sujets composés, du type *Daniel and me*
6. Parmi les 68 cas, 3 sont pour des propositions sans verbe conjugué (« non finies ») mais qui sont des propositions individuelles donc non déclassées.

8. REFERENCES

Consigny, A. 2005. 'Je ne suis plus moi-même : us et abus du pronom réfléchi anglais « myself » à partir d'un corpus tiré du web'. Communication aux Journées d'études sur la représentation linguistique de la personne, Rouen.

Fina, A.D. 1995. 'Pronominal choice, identity and solidarity in political discourse' dans *Text*, vol. 15/3, p. 379-410.

Francis, G. 1989. 'Thematic selection and distribution in written discourse' dans *Word*, vol. 40, p. 201-222.

Gomez-Gonzales, M.A. 1994. 'The relevance of theme in the textual organisation of BBC news reports' dans *Word*, vol. 45, p. 293-305.

Halliday, M.A.K. 1994. *An Introduction to functional grammar*. 2e éd. Londres : Arnold.

Halliday, M.A.K. et R. Hasan. 1976. *Cohesion in English*. Londres : Longman.

Halliday, M.A.K. et R. Hasan. 1989. *Language, context and text : aspects of meaning in a social-semiotic perspective*. Oxford : Presses Universitaires d'Oxford.

Herring, S.C. et al. 2004. 'Bridging the gap : A genre analysis of weblogs' dans *Proceedings of the 37th Hawaii International Conference on System Sciences* Los Alamitos : IEEE Press.

Hyland, K. 2003. 'Self-citation and self-reference : credibility and promotion in academic publication' dans *Journal of the American Society for Information Science and Technology*, vol. 54/3, p. 251-259.

Kjellmer, G. 1986. 'Us Anglos are a cut above the field : on objective pronouns in nominative contexts' dans *English Studies*, vol. 67, p. 445-449.

Kuo, C.H. 1998. 'The use of personal pronouns : role relationships in scientific journal articles' dans *ESP Journal*, vol. 18/2, p. 121-138.

Lüdeling, A., S. Evert et M Baroni. (à paraître). 'Using web data for linguistic purposes' dans Hundt, M., C. Biewer et N. Nesselhauf (eds.), *Corpus linguistics and the web*. Amsterdam : Rodopi.

Marlow, C. 2004. 'Audience, structure and authority in the weblog community'. Communication à la 54e conférence de l'International Communication Association.

Morley, B., A. Renouf et A. Kehoe. 2003. 'Linguistic research with XML/RDF-aware WebCorp tool' dans *Actes du colloque WWW03*, Budapest.

Owen, C. 1996. 'Do concordance require to be consulted?' dans *ELT Journal*, vol. 50, p. 219-224.

Quirk, R. et al. 1985. *A Comprehensive grammar of English*. Londres : Longman.

Renouf, A. 2004. 'Investigating phrasal productivity in web-based text'. Communication au séminaire JASGIL 04, Strasbourg. A paraître dans *RANAM*.

Sharoff, S. (en soumission) 'Open-source corpora: using the net to fish for linguistic data'.

Smyk, D. 2005. 'Computer-mediated communication and language change : a case of weblog'. Communication au séminaire JASGIL 05, Freiburg im Brisgau.

Squire, J.D. 2003 'Blogging : Digital Davids in an Internet Goliath' dans *SLIS New*, Université d'Indiana. www.slis.indiana.edu/news

Thompson, G. 2004. *Introducing functional grammar*. 2e éd. Londres: Arnold.

Thompson, G. et P. Thetela. 1995. 'The sound of one hand clapping'. dans *Text*, vol. 15/1, p. 103-127.

Wales, K. 1996. *Personal pronouns in present-day English*. Cambridge : Presses Universitaires de Cambridge.

Watson-Todd, R. 2001. 'Induction from self-selected concordances and self-correction' dans *System*, vol. 29, p. 91-102.

UTILISATION DU WEB COMME MEGA-BASE DE DONNÉES LINGUISTIQUE : APPLICATION À LA TRADUCTION DE COOCCURRENCES LEXICALES FRANÇAISES-ANGLAISES

Chrystel Millon & Stéphanie Léon
Equipe DELIC – Université de Provence
29, Av. Robert Schuman – 13621 Aix-en-Provence Cedex 1
fanny.leon@orange.fr, Chrystel.Millon@up.univ-mrs.fr

1. INTRODUCTION : PROBLÉMATIQUES DE TRADUCTION AUTOMATIQUE

Bien qu'elle ait été la première application non-numérique de l'informatique, la traduction automatique a connu des débuts décevants, qui ont jeté un discrédit sur cette technologie pendant plusieurs décennies. Toutefois, des progrès ont été accomplis au cours des dernières années, en raison de l'avènement du Web dans un contexte fortement multilingue, de l'accroissement très important de la couverture des dictionnaires présents dans les systèmes de traduction, et de la prise en compte d'un nombre croissant d'expressions composées. Par exemple, le système Systran[1] traduit désormais correctement du français vers l'anglais des expressions telles que :

vol à main armée → *armed robbery*

vol à la roulotte → *stealing from parked vehicles*

En revanche, dès que l'on sort de ces listes d'expressions figées, on retombe rapidement dans des erreurs de traduction qui gênent considérablement la compréhension, et lui donnent même parfois un caractère surréaliste. Ainsi, Systran utilise la traduction la plus fréquente du mot *vol*, c'est-à-dire *flight* (VOL AERIEN), dans toutes les autres situations. Si *réserver un vol* est correctement traduit (*to reserve a flight*), *commettre un vol* est traduit par *to make a flight*, ce qui est totalement incompréhensible dans ce contexte pour un anglophone.

Pourtant, la cooccurrence lexicale[2] *commettre-vol* est un indice désambiguïsateur très fort, qui, si elle était correctement traduite dans une base de données (*to commit-theft*), pourrait servir à générer des traductions correctes. La combinatoire est toutefois beaucoup plus ouverte qu'avec les expressions composées mentionnées plus haut et la constitution manuelle d'une base de données de combinaisons lexicales à grande échelle est une tâche à peu près impossible. Les dictionnaires bilingues se contentent d'ailleurs de rares indications ponctuelles, se fiant au jugement du lecteur et à sa connaissance du monde, que l'on ne peut guère espérer d'une machine.

Afin de constituer de telles bases de données bilingues, de nombreux travaux se sont appuyés sur des corpus parallèles ou alignés[3] (Véronis, 2000). Toutefois, ces méthodes présentent diverses limites : de telles ressources sont peu nombreuses et concernent des

domaines restreints. Le paysage est un peu plus souriant avec des corpus comparables[4] (Morin *et al.*, 2004), mais les contraintes restent fortes.

Le Web, qui génère des besoins considérables de traduction, offre en même temps un réservoir gigantesque de données qui peuvent être exploitées par des moyens automatiques, en particulier grâce à des moteurs de recherche tels que Google ou Altavista. En se basant sur les travaux de (Kilgarriff & Grefenstette, 2003), on peut estimer à environ 100 milliards le nombre de mots indexés par Google pour la seule langue anglaise. Même si les données du Web sont moins contrôlées, et donc plus « bruitées », elles permettent d'envisager un changement radical d'échelle pour les méthodes basées sur les données, à condition de développer des méthodes et des techniques adaptées. Depuis quelques années, divers domaines du TAL utilisent le Web en tant que ressource de données linguistique. Cao & Li (2002) proposent une méthode automatique de génération et de sélection de traductions pour les syntagmes nominaux de l'anglais vers le chinois à partir du Web. Les résultats montrent que le Web peut être utile tant pour l'acquisition de données que pour une aide à la validation.

L'objectif de notre article est de proposer une méthodologie de validation semi-automatique de traductions anglaises via le Web, à partir de cooccurrences lexicales françaises du type *NOM ADJECTIF*, *NOM1 DE NOM2* et *VERBE NOM*(*objet*) extraites d'un corpus français de pages Web en langue générale. Pour revenir à nos exemples *commettre un vol* et *réserver un vol*, Google nous permet de valider les traductions correctes, grâce à leur nombre d'occurrences. Par exemple, la requête [*″commit a flight″* OR *″commit the flight″*] retourne seulement 13 résultats. La requête [*″commit a theft″* OR *″commit the theft″*] retourne quant à elle 5 110 résultats. Parmi ces deux traductions candidates, les résultats sélectionnent de façon écrasante la cooccurrence lexicale satisfaisante (*to commit-theft*).

2. METHODOLOGIE ET TRAITEMENT DES DONNEES

2.1. Extraction automatique des cooccurrences lexicales françaises

Notre méthode a été testée sur la combinatoire lexicale de 10 noms français très polysémiques (*barrage*, *détention*, *formation*, *lancement*, *organe*, *passage*, *restauration*, *solution*, *station* et *vol)*. Ces mots ont été jugés comme les plus polysémiques parmi 200 noms de fréquence équivalente, lors du projet Senseval (Véronis, 1998) et constituent donc un banc de test difficile. Nous exploitons la version lemmatisée et étiquetée morpho-syntaxiquement du corpus de pages Web francophones élaboré par Jean Véronis (Véronis, 2003) autour de ces 10 noms.

L'acquisition automatique des cooccurrences lexicales françaises de type *NOM ADJECTIF*, *NOM1 DE NOM2*[5] et *VERBE NOM*(*objet*) utilise une version améliorée du programme employé dans (Millon, 2004), dans lequel des filtres linguistiques d'extraction sont utilisés (représentation de patrons syntaxiques, filtres de candidats indésirables). Les cooccurrences lexicales sont ensuite soumises à un seuil limite de fréquence fixé à au moins 10 occurrences. En effet, un nombre important de cooccurrences moins fréquentes est « accidentelle » (comme par exemple *barrage violet*). L'objectif est d'extraire une majorité de cooccurrences lexicales « propres » du côté français. À l'inverse, des cooccurrences lexicales caractéristiques sont perdues (ainsi *barrage hydraulique* a une fréquence de 4).

2.2. Génération automatique des traductions potentielles

Pour chacune des cooccurrences lexicales françaises non présentes dans le *Collins Pocket French-English Dictionary*, nous générons automatiquement toutes les traductions possibles via le *Collins Pocket*. Reprenons pour exemple *réserver-vol*. Le *Collins Pocket* donne les traductions suivantes pour les unités lexicales sources *vol* et *réserver* (unité lexicale source vers unité lexicale cible) :

vol → *flight, theft, flying*

réserver → *to reserve, to book*

Notre programme génère toute la combinatoire :

réserver un vol → *to reserve a flight, to reserve a theft, to reserve a flying, to book a flight, to book a theft, to book a flying*

Le Tableau 1 donne la quantité de traductions potentielles générées.

	CLs françaises après filtre	Traductions générées	Moyenne par CL française
N ADJ	113	1215	11
N DE N	173	5155	30
V N	57	1012	18
TOTAL	**286**	**7382**	**19**

Tableau 1 : Résultats de l'étape de génération des traductions candidates

2.3. Interrogation automatique de Google

Le moteur de recherche Google a été interrogé automatiquement à l'aide de l'interface de programmation d'applications API (*Application Programming Interface*)[6] afin de récupérer le nombre d'occurrences[7] de chaque cooccurrence lexicale anglaise candidate. Ces fréquences seront utilisées lors de la validation[8].

Pour chaque traduction potentielle, nous générons un ensemble de requêtes (voir Tableau 2), en considérant les mots de la requête comme une expression exacte, via l'utilisation des guillemets. La recherche est restreinte aux pages Web de langue anglaise. La combinaison booléenne pour les patrons syntaxiques de type *NOM ADJECTIF* et *VERBE NOM*(*objet*) ramène un ensemble de résultats qui prend en compte les variations dues aux changements d'article, comme dans l'exemple « *commit a theft* » *OR* « *commit the theft* ».

L'utilisation d'articles dans les requêtes du patron syntaxique *NOM ADJECTIF* permet également de réduire le problème de l'ambiguïté catégorielle.

Patron syntaxique source	Requête (en langue cible)
NOM ADJECTIF	"the ADJ NOM" OR "a ADJ NOM"
VERBE NOM(objet)	"VERBE the NOM" OR "VERBE a NOM"
NOM1 de NOM2	"NOM_1 of NOM_2"
	"NOM_2 NOM_1"

Tableau 2 : Patrons des requêtes des cooccurrences lexicales anglaises

Le patron syntaxique *NOM1 DE NOM2* pouvant être traduit par différentes structures en anglais selon la cooccurrence sémantique considérée entre les deux objets, nous traitons séparément deux types de structures dans les requêtes (Chuquet & Paillard, 1987) : d'une

part, le patron *N2 N1* marque une cooccurrence étroite entre les deux noms et d'autre part, la structure *N1 of N2* accorde la priorité à l'élément repéré (*N2*)[9].

2.4. Validation semi-automatique des traductions potentielles anglaises

2.4.1. FILTRE AUTOMATIQUE Un filtre simple a été appliqué aux traductions restantes. Nous ne conservons que celles dont la fréquence sur le Web est au moins égale à un millième des occurrences du mot cible. Prenons pour exemple réserver-vol et deux de ses traductions candidates book a/the flight et book a/the theft :

Seuil_theft : 2150000 / 1000 = 2150

Seuil_flight : 5760000 / 1000 = 5760

La cooccurrence lexicale *book a/the flight* (avec une fréquence de 244 000, donc supérieure au seuil limite pour le nom cible *flight*) est retenue, tandis que *book a/the theft* (avec une fréquence de 61) est rejetée.

Ce filtre provoque parfois des cas de silence. À partir de la cooccurrence lexicale *barrage hydro-électrique*, la traduction *hydroelectric dam* est retenue (fréquence de 3 920 et seuil à 1 910), tandis que *hydroelectric barrage*, également valide, a une fréquence de 32 (pour un seuil à 139) et est rejetée. Notre approche favorise la précision, car il s'agit de compléter le plus automatiquement possible des ressources existantes. L'augmentation du bruit obligerait à un filtrage manuel des résultats plus long et coûteux. Après le filtre automatique, 7,5 % des cooccurrences anglaises sont conservées (Tableau 3).

2.4.2. VALIDATION MANUELLE Une validation manuelle nous permet d'évaluer les cooccurrences lexicales restantes, en faisant appel à divers dictionnaires de langue, ainsi qu'à nos connaissances. Les traductions candidates les plus « délicates » sont vérifiées en contexte sur le Web, par reformulation de la requête, ainsi que soumises au jugement de locuteurs.

	Traductions générées	***Filtre automatique*** **Filtre seuil fréquence**		***Validation manuelle*** **Traductions valides**	
N ADJ	1215	136	11,2 %	132	97,1 %
N DE N	5155	351	6,8 %	270	76,9 %
V N	1012	63	6,2 %	56	88,9 %
TOTAL	**7382**	**550**	**7,5 %**	**458**	**83,3 %**

Tableau 3 : Résultats de la validation des traductions

3. PREMIERS RESULTATS

La précision globale est de 83,3 %. La méthode fonctionne particulièrement bien pour les patrons syntaxiques *NOM ADJECTIF* (97,1 %) et *VERBE NOM*(*objet*) (88,9 %) (Tableau 4). Le patron *NOM1 DE NOM2* pose davantage de difficultés en traduction (76,9 %).

MOT	**N ADJ**		**N de N**		**V N**		**TOTAL**	
SOURCE	***Nb total***	***% valide***	***Nb total***	***% valide***	***Nb total***	***% valide***	***Nb total***	***% valide***
TOTAL	**136**	**97,1**	**351**	**76,9**	**63**	**88,9**	**550**	**83,3**

Tableau 4 : Précision globale des cooccurrences lexicales anglaises évaluées

Le Tableau 5 présente le nombre moyen de traductions valides par cooccurrence lexicale française.

	CLs françaises	Traductions validées	Moyenne par CL
N ADJ	63	132	2,1
N de N	124	270	2,2
V N	31	56	1,8
MOYENN	**218**	**458**	**2,1**

Tableau 5 : Nombre moyen de traductions valides par cooccurrence lexicale française

Le Tableau 6 présente un exemple de traductions pour la cooccurrence lexicale *construire-barrage*.

PATRON	CL FRANCAISE	TRADUCTION
V N	construire-barrage	to build-barrage
		to build-barricade
		to build-dam
		to build-roadblock
		to construct-dam
		to erect-barricade
		to erect-roadblock

Tableau 6 : Traductions des cooccurrences lexicales de barrage

4. ANALYSE DES PROBLEMES ET PERSPECTIVES

4.1. Erreurs et difficultés

4.1.1. ERREURS SYNTAXIQUES Certaines erreurs sont dues à des problèmes d'ambiguïté de rattachement syntaxique concernant les traductions candidates testées sur l'API Google. Le moteur de recherche ne permettant pas un accès direct aux catégories morpho-syntaxiques, une analyse syntaxique des contextes des traductions candidates n'a pu être réalisée. Prenons l'exemple suivant :

The Library of Congress ***set the changeover*** *date.*

La cooccurrence lexicale recherchée est *set the changeover*. Or dans ce cas, *changeover* est régi par le nom *date* et non par la forme verbale *set*. Une autre limite concerne les problèmes d'ambiguïté catégorielle. Les formes en *-ing*, par exemple, sont propices à ce type d'erreurs. Une perspective d'amélioration des problèmes d'ordre syntaxique est donc manifestement l'application d'un analyseur syntaxique aux pages Web anglaises.

Enfin, il arrive qu'une cooccurrence lexicale ne soit pas traduite par une structure de même longueur. Par exemple, *barrage routier* se traduit par une unité lexicale simple : *roadblock*. La formation des mots composés en anglais, avec ou sans espace ou tiret, est évidemment un cas extrêmement difficile, mais on peut envisager de générer des requêtes du type N-N ou NN (sans espace).

4.1.2. ERREURS SEMANTIQUES Un type d'erreurs d'ordre sémantique concerne l'acquisition de cooccurrences lexicales anglaises valides, mais non correspondantes à la cooccurrence lexicale source, comme dans l'exemple :

cours de formation --> *group rate* (59900 occurrences)

Ici, group rate signifie tarif de groupe.

De plus, la traduction d'une cooccurrence lexicale n'est pas toujours obtenue de façon compositionnelle (Melamed, 2001, cité par Morin *et al*., 2004). Par exemple, en anglais, il n'existe pas une traduction littérale de *forcer un barrage* : la traduction va dépendre du contexte situationnel (*to drive through a roadblock, to run through a roadblock*, etc.*)*.

Une autre limite est due à l'absence de certains usages dans le dictionnaire. C'est le cas par exemple de l'acception sportive du nom *barrage* (*playoff*). Ainsi, la cooccurrence lexicale *match de barrage* est traduite par *weir game* et *barrage game* au lieu de la traduction correcte *playoff game*.

4.1.3. ERREURS TECHNIQUES Google ne prend pas en compte des phénomènes tels que la ponctuation, les majuscules, ou la marque du génitif ('s), lors des recherches sur le Web. Certaines cooccurrences lexicales anglaises erronées, comme to reserve-theft (fréquence de 3) n'ont pas une fréquence nulle pour cette raison. L'exemple suivant montre que les mots reserve et a theft appartiennent à deux syntagmes différents :

> *A man will face court next month charged with stealing three date palms from a Swansea* ***reserve, a theft*** *which sparked three months of community outrage.*

Le filtre sur les fréquences permet d'éliminer une partie des cooccurrences lexicales erronées. Néanmoins, ces problèmes concernent également des traductions correctes, et « bruitent » quelque peu les fréquences de l'API Google[10].

4.2. Perspectives d'amélioration du protocole

4.2.1. CHANGEMENT DE SEUIL DE FILTRAGE DES COOCCURRENCES LEXICALES FRANÇAISES Les cooccurrences lexicales françaises qui comptent moins de 10 occurrences au sein du corpus sont éliminées. Or, un certain nombre de cooccurrences lexicales correctes ont des fréquences inférieures à ce seuil. C'est le cas, par exemple pour l'usage sportif de barrage (match de barrage, barrage aller, barrage retour, etc.). L'algorithme HyperLex (Véronis, 2003 & 2004) nous permettrait d'identifier les usages peu fréquents des mots (jusqu'à environ 1 % des occurrences). Une amélioration consisterait à ajuster le seuil des cooccurrences lexicales françaises selon la fréquence de l'usage du nom concerné.

4.2.2. DESCRIPTION DES PATRONS SYNTAXIQUES DE L'ANGLAIS Contrairement à notre méthode d'extraction des cooccurrences lexicales françaises, ne sont pris en compte que les patrons de « base » de l'anglais sans autres variations que celles de l'article dans le cadre de notre acquisition des traductions. Ces patrons « de base » ont donné un premier éventail de résultats qui ont permis d'évaluer notre méthode. Des améliorations ultérieures visent à augmenter le panel de patrons morpho-syntaxiques anglais des traductions candidates, en procédant à une analyse syntaxique du contenu des pages Web prospectées.

4.2.3. VARIATIONS MORPHO-SYNTAXIQUES AU SEIN DES REQUETES ANGLAISES Une amélioration vise à considérer des variations morpho-syntaxiques au sein de nos requêtes telles que des variations dues aux formes verbales ou au pluriel, comme dans l'exemple :

> *"commit a theft", "commit the theft", "commit thefts" , "commit the thefts"*, etc.

5. CONCLUSION

Nous avons décrit une méthode d'acquisition semi-automatique de cooccurrences de traductions du français vers l'anglais, en montrant que le Web s'avère être un outil efficace

d'aide à la validation de traductions candidates. Les résultats sont intéressants pour les patrons syntaxiques de type *NOM ADJECTIF* (précision de 97,1 %) et *VERBE NOM*(*objet*) (précision de 88,9 %). La méthode reste imparfaite pour le patron *NOM1 DE NOM2*, mais le taux de précision est honorable (76,9 %), surtout étant donné la difficulté volontaire du banc de test choisi (mots très polysémiques). Ces résultats constituent un premier échantillon significatif des possibilités qu'offre le Web en matière de validation de cooccurrences lexicales. Nos perspectives d'évolution concernent la mise en place d'une analyse syntaxique des cooccurrences candidates en contexte, ainsi qu'une désambiguïsation des cooccurrences lexicales à traduire.

6. RÉFÉRENCES

Benson M. 1990. 'Collocations and general-purpose dictionaries', *International Journal of Lexicography*, vol. 3(1), p. 23-35.

Cao Y., Li H. 2002. 'Base noun phrase translation using web data and the EM algorithm' in *Proceedings of CoLing*.

Chuquet H., Paillard M. 1987. *Approche linguistique des problèmes de traduction: anglais-français*, Gap, Ophrys.

Cruse D. A. 1986. *Lexical Semantics*, Cambridge, Cambridge University Press.

Katz J. J, Fodor J. A. 1964. 'The structure of a semantic theory' in Fodor J. A. and Katz J. J. (eds.), *The Structure of Language*, chapter 19, p. 479-518.

Kilgarriff A., Grefenstette G. 2003. 'Introduction to the Special Issue on the Web as Corpus', *Computational Linguistics*, *29*(3), p. 333-348.

Melamed I. D. 2001. *Empirical methods for exploiting parallel texts*, MIT Press.

Millon C. 2004. 'Acquisition de cooccurrences lexicales désambiguïsées à partir du Web', *Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2004)*, Fès, (Maroc).

Morin E., Dufour-Kowalski S. & Daille B. 2004. 'Extraction de terminologies bilingues à partir de corpus', in *Actes de TALN'2004*, Fès (Maroc).

Smadja F. 1993. 'Retrieving collocations from text : Xtract', *Computational Linguistics*, Vol. 19, p. 143-177.

Véronis J. 1998. 'A study of polysemy judgements and inter-annotator agreement', *Programme and advanced papers of the Senseval workshop*, England : Herstmonceux Castle, p. 2-4.

Véronis J. (Ed.) 2000. *Parallel Text Processing : Alignment and use of translation corpora*, Kluwer Academic Publishers.

Véronis J. 2003. 'Hyperlex : cartographie lexicale pour la recherche d'informations' in *Actes de TALN'2003*, Batz-sur-mer (France) : ATALA, p. 265-274.

Véronis J. 2004. *HyperLex : cartographie lexicale pour la recherche d'informations*. Rapport Interne Equipe DELIC, Université de Provence. (En ligne : http://www.up.univmrs.fr/veronis/pdf/2004-hyperlex-rapport.pdf)

Wilks Y. A. 1975. 'Preference Semantics' in Keenan E. (ed.), *The Formal Semantics of Natural Language*, Cambridge University Press, p. 329-348.

7. NOTES

1. http://www.systransoft.com/

2. La littérature présente une terminologie disparate. Certains parlent de « préférences lexicales » (Wilks, 1975), de « restrictions de sélection » (Katz & Fodor, 1964) ou encore de « collocations » (Benson, 1990 ; Smadja, 1993 ; Cruse, 1986). Nous employons le terme « cooccurrence lexicale », plus neutre, défini comme une relation lexicale entre deux lexèmes liés syntaxiquement.

3. Ensemble de textes alignés avec leur traduction au niveau du paragraphe, de la phrase, des expressions ou des mots.
4. Corpus de langues différentes traitant du même domaine mais non parallèles.
5. Le nom source peut se trouver en position NOM1 ou NOM2.
6. http://www.google.com/apis/
7. Des différences ont été remarquées entre le nombre de résultats renvoyés par l'API et par l'interface Web.
8. Si les fréquences de Google sont peu fiables dans le cadre de certaines configurations de requêtes dans « tout le Web » (http://aixtal.blogspot.com/2005/02/web-le-mystre-des-pages-manquantes-de.html), ce problème ne concerne pas l'utilisation que nous faisons de Google puisque nous limitons les requêtes à une langue donnée.
9. Le cas du génitif (N1's N2) n'est pas pris en compte dans le cadre de cette étude.
10. Une autre limite de l'API Google est de ne pouvoir lancer que 1 000 requêtes par jour
.

CONTRIBUTION DE LA LEXICOMETRIE À L'APPROCHE SÉMANTIQUE DES CORPUS. LA FORME « TEXTE » DANS UN CORPUS DES « ÉTUDES LITTÉRAIRES ».

Driss Ablali
Université de Franche-Comté

1-INTRODUCTION

Ce travail s'inscrit dans le cadre d'une étude sémantique des concepts. L'objectif est de montrer que les concepts dans le discours scientifique peuvent être étudiés comme des thèmes. Car l'avantage que nous avons en sciences humaines, c'est d'avoir des théories qui s'écrivent dans des textes, contrairement à certaines sciences dites exactes. Il y a différentes façons pour donner forme à une théorie : formules mathématiques, formules chimiques, écriture, etc. Dans les sciences humaines, théorie et écriture sont inséparables. La théorie naît dans l'écriture, dans les textes et dans les discours. Le théoricien construit ses concepts et ses notions en écrivant des articles, des livres, en donnant des conférences, en assurant des cours. Il est donc producteur de textes. Mais cette textualité a souvent été ignorée, sauf dans le cadre de la terminologie ou de la didactique. D'où les questions suivantes : peut-on analyser le discours théorique comme on analyse des textes littéraires, journalistiques ou politiques ? En quoi le discours d'un scientifique diffère-t-il de celui d'un écrivain, d'un journaliste ou d'un politicien ? Les concepts scientifiques peuvent-ils être étudiés comme des formes textuelles ? Les concepts de « l'autobiographie », de l'« intertextualité » ou du « dialogisme », par exemple, peuvent-ils être traités comme les thèmes du paysage chez Bonnefoy, de l'immigration dans un quotidien comme *Le Figaro*, ou de l'écologie dans le discours des verts ? C'est à partir de semblables réflexions que ce travail est né sur la textualité du discours scientifique. Il est né aussi autour d'une question unifiante, de pleine actualité dans les théories textuelles : celle du genre.

2. LA QUESTION DES GENRES

Pour commencer, quelques remarque sur l'expression « études littéraires ». Nous employons cette expression pour désigner des travaux qui viennent des lettres ou de la linguistique, et non pas pour désigner des études qui portent exclusivement sur des textes littéraires. Il s'agit d'un travail qui se propose d'analyser un texte, et nous prenons le texte, au sens sémiotique, comme tout objet soumis à l'analyse, peu importe le médium qu'il emprunte : littérature, cinéma, peinture, musique, etc. Et puisque les genres sont indexés sur des pratiques sociales, on entendra par « études littéraires » ce qui est situé dans une tradition universitaire, voire dans un cadre éditorial et académique donné, à savoir *l'article universitaire*. On a exploité autant que possible cet empan, celui du genre et en restant dans le discours universitaire avec une pertinence fondée ici sur de fortes unités : la *langue*, que

du français ; *l'unité de domaine*, le discours universitaire ; le *support*, l'article ; la *datation*, à partir des années 70.

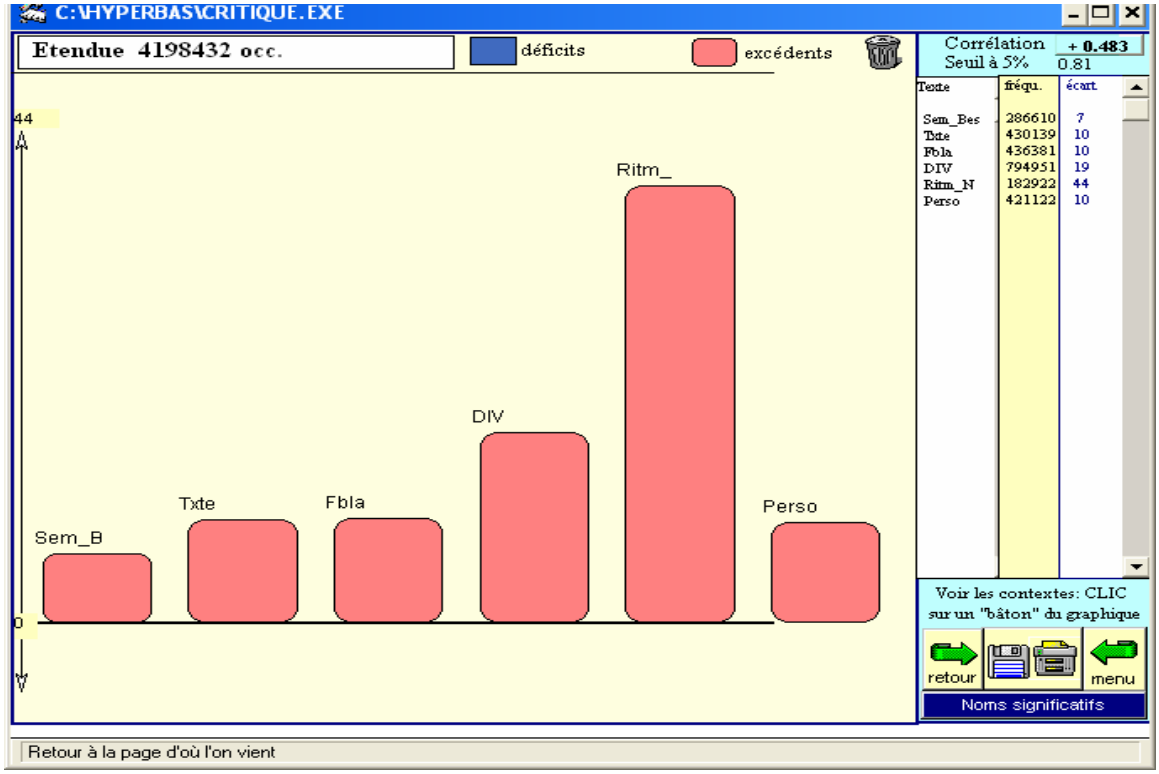
3. LE CORPUS :

Notre corpus électronique avoisine les : 4 198 432 mots. Il comprend des articles de revues françaises et francophones, toutes universitaires, et sur support électronique. Les textes du corpus sont assez homogènes au niveau de la taille, à part *RITM* qui domine tous les autres. Il est aisé de constater que la taille des différentes sections varie beaucoup. Certaines sont d'une étendue importante tandis que d'autres sont très courtes Il est vrai que ces différences de taille pourraient nuire dans certains traitements statistiques et occasionner des erreurs d'interprétation. Mais cette exception n'a pas d'influence particulière sur le traitement que nous voulons en faire, dans la mesure où l'accent est moins mis sur la distance lexicale ou sur l'évolution du vocabulaire que sur la construction des concepts scientifiques dans le discours des études littéraires. Le tableau ci-dessous rend compte des caractéristiques de notre corpus :

Revue	***Lieu de publication***	***Nombre d'occurrences***
SEMEN	Université de Franche-Comté	286 610
RITM	Université de Paris-X-Nanterre	1 829 229
TEXTE	Université de Toronto	430 139
FABULA	Site universitaire	436 681
DIVERS	Fichier qui regroupe plusieurs revues toutes universitaires	794 951
PERSO	Articles demandés directement aux auteurs, ou collectés dans des sites personnels	421 122

Tableau 1. Corpus

PS : Nous voulons rappeler que dans ces revues nous n'avons pris que les articles qui sont conformes à la définition que nous avons donnée ci-dessus des « études littéraires ». À cet égard, vu la taille de ce corpus, on pense que l'on peut le considérer comme un corpus de référence, un échantillon de la population des études littéraires, sans qu'il soit menacé ni par « random error » ou « bias error », pour parler comme Biber, (1993 : 243)**.** Le graphique ci-dessous, réalisé avec *Hyperbase*, permet de visualiser ces variations :



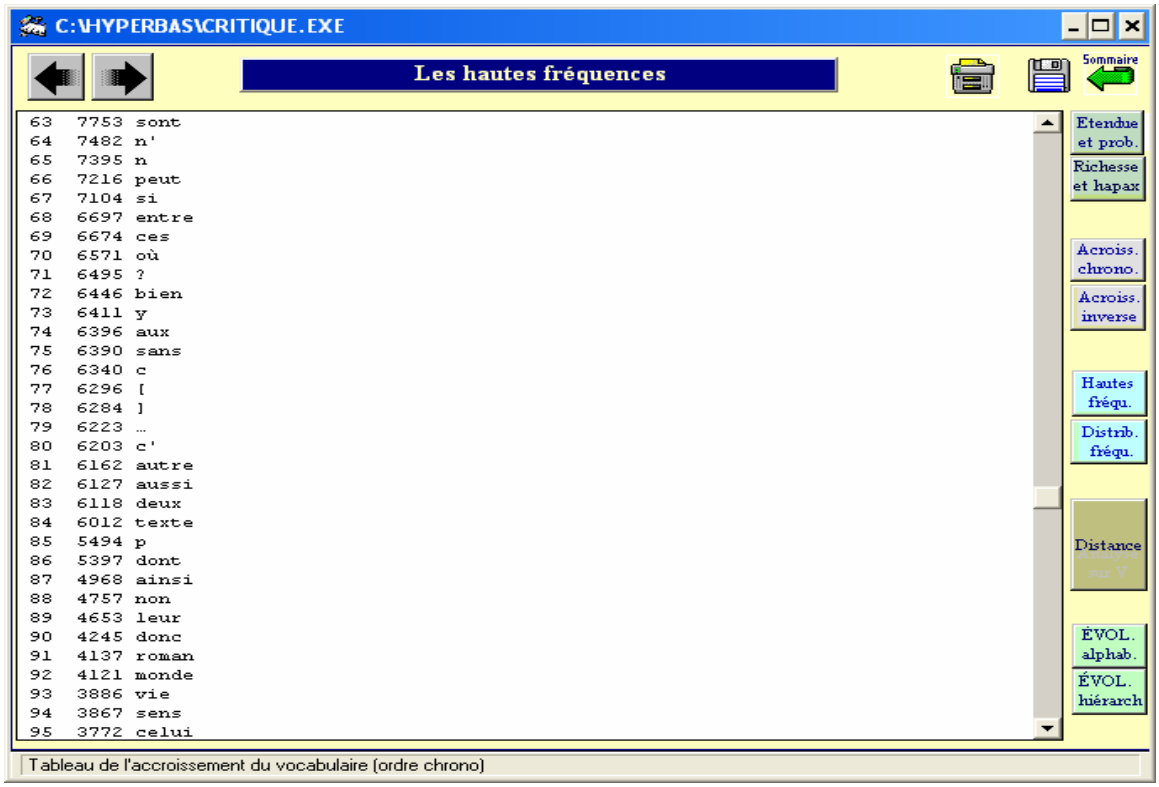
Graphique 1. Étendue relative du corpus « Etudes Littéraires ».

4. LA FORME « TEXTE », CONCEPT OU SIMPLE FORME ?

Dans cette communication, il s'agit de montrer, au moyen du logiciel *Hyperbase*, que les « formes » dans le discours scientifique peuvent être appréhendées comme le sont les thèmes dans le discours littéraire, journalistique ou politique. Et pour le faire de façon impartiale, il s'avère que les techniques de lexicométrie sont parfaitement adaptées comme plate-forme pour ces analyses.

À partir de la réflexion de F. Rastier, « une forme n'est pas une unité discrète, stable, identique à elle-même : les formes (que l'on opposait faussement à des substances) ne sont que des figures, qui contrastent sur des fonds » (2003 : 100), on s'attachera à voir la distribution d'une forme précise en tenant compte de son entourage lexical et morphosyntaxique pour souligner combien est arbitraire la frontière du mot et combien indispensable une typologie des contextes dans le cadre de la sémantique des concepts, car les formes ne prennent leur sens que dans une textualité en s'opérant de façon située dans un contexte donné.

La forme en question est la forme « texte ». On en dénombre plus de 6 000 occurrences. À quoi tient cette haute fréquence du *texte* dans le corpus ? Concept ou simple forme ? De quoi parle-t-on dans les études littéraires lorsqu'on évoque la question du *texte* ? Y a-t-il une activité définitoire dans l'utilisation de cette forme qui pourrait la distinguer des autres théories, sémiotique, linguistique textuelle ou psycholinguistique ? Quel est l'écart entre les revendications des études littéraires qui se veulent textuelles et le reflet qu'en donnent les bases de données ? Le graphique suivant rend compte de la distribution des hautes fréquences dans le corpus selon le principe de l'écart réduit calculé par le logiciel *Hyperbase* :



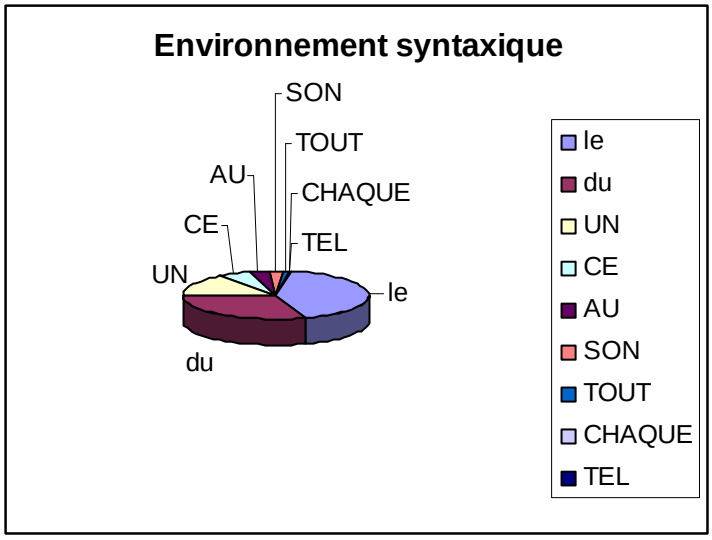
Graphique 2. Les hautes fréquences du corpus « études littéraires ».

5. ENVIRONNEMENT SYNTAXIQUE DES FORMES « TEXTE » ET « TEXTES ».

Pour commencer, nous avons opté pour la distribution de cette forme dans le corpus en regardant de plus près les formes qui la précèdent :

	Contexte gauche	Occurrences
La forme « TEXTE »	**Le**	**2 235**
	Du	**1 643**
	Un	**767**
	Ce	**300**
	Au	**194**
	Son	**123**
	Tout	**48**
	Chaque	**21**
	Tel	**10**
Total	**5 341**	

Tableau 2. Environnement syntaxique



Graphique 3. Environnement syntaxique de la forme « texte ».

Les autres emplois sont répartis entre des formes comme : « En texte », « À texte », « Même texte », « Entre texte » que nous n'avons pas jugés pertinents.

Le *texte* avec ses 6 012 occurrences est une lexie très importante dans notre corpus. Avant de commenter son entourage syntaxique, nous proposons que l'on regarde aussi la même forme, mais en changeant de nombre, *textes*. Y a-t-il une différence d'emploi entre les deux formes ? Ont-elles le même entourage syntaxique ? Y a-t-il un contexte propre à chaque forme ? Avant de répondre à ces questions, regardons la forme au pluriel.

La forme enregistre 1 969 occurrences, presque le tiers du singulier, son entourage syntaxique se répartit de la façon suivante :

	Contexte gauche	**Occurrences**
La forme « TEXTES »	**DES**	**513**
	LES	**512**
	DE	**212**
	CES	**150**
	SES	**98**
	DEUX	**75**
	AUX	**38**
	CERTAINS	**25**
	AVANT	**34**
	GRANDS	**17**
Total	**1 674**	

Tableau 3. Environnement syntaxique

En regardant de plus près les contextes gauches de la forme *texte*, aussi bien au singulier qu'au pluriel, on se rend compte que dans la plupart des cas ils participent à l'expression de l'intertextualité et de l'interprétation. En d'autres termes, c'est le texte en tant qu'objet d'analyse dont il est question, le texte comme objet des études littéraires, d'où les exemples suivants :

Pour **le texte** donné à la Revue de la langue française

Le texte de 1839 accentue à la fin l'aspect personnel par un " nous " d'auteur

Le texte de 1837 est beaucoup plus court que les suivants

Dans **le texte** de Georges Bataille

Il est intéressant de remarquer à ce titre que **le texte** de L'Impossible déroute totalement la cohérence

Il est dans **le texte** de Chraïbi des données qui disposent le personnage de la mère à assumer le rôle de l'écrivat

Le texte fournit un éclairage en retour sur les événements ignorés ou énigmatiques

Ainsi, notre observation sur la nécessité du retour à une réflexion sur le lien entre **le texte** et

Tout se passe comme si **le texte** des Carnets était chargé de faire un voyage à la place de son auteur, objet qu'il représente

Cette disposition graphique configure rythmiquement **le texte** et contribue à sa dramatisation lyrique

Le parallélisme embraye **le texte** sur l'extra-textuel en mettant en corrélation deux programmes narratifs

Tous les parcours effectués jusqu'à présent étaient encodés, programmés dans **le texte**

Pour le pluriel :

Mais surtout ce sont **les textes** des deux séries qui présentent des éléments communs non seulement diégétiques mais littéraux

Presque tous les livres me tombent des mains, hormis **les textes** religieux

Cette nouvelle donne implique qu'on analyse les interactions, y compris dans **les textes** littéraires

L'analyse conduira ensuite à souligner comment un déplacement progressif va s'opérer, lisible dans l'importance croissante que les écrivains africains vont porter à la question du savoir que **les textes** oraux sont susceptibles de véhiculer

C'est le cas en particulier, dans les **textes narratifs**, lorsque le narrateur interrompt le fil du récit pour mettre en scène un personnage de conteur ou de chanteur qui va présenter un texte

On comprendra par là qu il est donc nécessaire de lire avec prudence **les textes** africains ou antillais

Il n'existe pas encore d'ouvrage fondamental sur **les textes** courts de fiction écrits par l'auteur

Ces écrits de Desnos devancent **les textes** que le Mexique inspirera à Antonin Artaud

La critique a établi dans **les textes** de Desnos une liaison entre l'amour et l'aventure maritime

Hoffmann fut également musicologue et critique musical : dans **les textes** qu'il écrivit en tant que tel

À partir de ces exemples, nous avons remarqué que la forme « texte » est indifférente à son entourage syntaxique. En d'autres termes, que ce soit au pluriel ou au singulier, dans une forme définie ou indéfinie, précédé d'un déterminant ou non, le sens est toujours le même. La forme renvoie à un phénomène intertextuel, le texte comme objet des études littéraires : le texte que l'on analyse, que l'on critique, que l'on interprète**,** mais jamais le

texte en tant qu'objet sémiotique. Il y a une différence, par exemple, pour la forme « sens » dans un contexte où il est question du sens en tant que notion métalinguistique, et le sens comme ce qui se dégage à la lecture d'une phrase ou d'un énoncé, la même chose pour « contexte » en tant concept, et « contexte » en tant élément constitutif d'une phrase ou d'un énoncé. Or dans notre corpus, cette opposition n'existe pas. Le « texte » en tant qu'objet métalinguistique n'enregistre aucune occurrence. Cela veut dire qu'il n'y a pas de réflexion conceptuelle autour de la forme « texte ». Toujours utilisée en relation directe avec un point de vue, la fiction, le mythe, la lecture, la réception, l'écriture, mais à aucun moment nous n'avons rencontré un contexte qui touche le genre en entier, c'est-à-dire les études littéraires. En d'autres termes, au sein d'un genre, ici les études littéraires, on ne théorise pas, et pourtant c'est la forme qui enregistre la plus grande fréquence. C'est ce que l'analyse thématique nous confirmera à travers les sèmes que cette forme reprend pour constituer un fond.

6. ENVIRONNEMENT THEMATIQUE

Le logiciel *Hyperbase* propose avec la fonction « Thème » de mesurer « l'attraction d'un mot pôle sur ses co-occurrents » et donne une représentation quantifiée du contexte d'une forme lexicale. *Hyperbase* offre ainsi un nouveau moyen, dans l'esprit des isotopies, purement documentaire, d'étudier les associations de mots s'appuyant sur la récurrence des mots qui l'entourent. L'indice de corrélation, constitué selon un test d'écart réduit à partir du volume d'occurrences d'un co-occurrent rencontré rapporté à son volume total et à la taille du sous corpus ainsi constitué, permet de mesurer cette attraction. L'importance du nombre de lexies fortement corrélées à un mot pôle est une indication précieuse de la pertinence du thème constitué. 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. En effet, il s'agit de répondre à des questions simples » (E. Brunet). Dans notre corpus, quels sont les mots qui sont employés ensemble – on supposera qu'ils appartiennent à un même univers ou une même structure ? Par exemple, quand la lexie « texte » est évoquée, quelles sont les autres formes que la lexie attire vers elle ?

Considérons *texte* comme un thème des études littéraires, non sans oublier de rappeler que le thème, d'après la sémantique interprétative, est une structure stable de traits sémantiques (ou sèmes), récurrente dans un corpus, et susceptible de lexicalisations diverses. Nous le ferons en prenant pour base une hiérarchie endogène de vocables, avec l'aide du logiciel *Hyperbase*, qui permet en effet d'effectuer une recherche de contextes automatique où chaque occurrence d'une forme est montrée dans son contexte (ici le paragraphe). Les mots qui se trouvent en tête de liste n'étonneront aucun lecteur. Voici ci-dessous quelques-uns des 6 012 contextes obtenus pour *texte*. Commençons par le singulier :

Ecart	corpus	extrait	mot
181.60	6012	3691	TEXTE
40.37	3130	721	LECTEUR
35.69	2059	506	LECTURE
30.22	3076	580	FICTION
22.64	28163	2577	D
22.61	85148	6634	LE
22.28	46995	3943	L
22.07	44210	3732	DU
19.84	379	114	FICTIONNEL
17.56	463	117	TRADUCTION
17.25	269	83	RÉFÉRENTIEL
16.91	78	40	
FICTIONALITÉ			
16.15	47	29	BALZACIEN
15.93	379	96	RÉCEPTION
15.87	343	90	TEXTUELLE
15.10	242	70	TEXTUEL
14.73	1969	272	TEXTES
14.66	843	151	RÉFÉRENCE
13.92	46967	3512	UN
13.87	198	58	NARRATAIRE
13.78	93	37	PARATEXTE
13.14	2072	265	LITTÉRAIRE
13.09	105	38	TEXTUELS
13.00	92	35	CIBLE
12.78	214	57	SIMON
12.69	157	47	
INTERTEXTUAL			
12.43	113	38	TEXTUELLES
11.76	373	76	COMMENTAIRE
11.72	82	30	BOUCHARD
11.47	37129	2735	QUE
10.88	118	35	INTERTEXTE
10.83	3144	331	AUTEUR
10.80	936	134	LIRE
10.64	3867	387	SENS
10.64	448	80	PEREC
10.22	349	66	FICTIF
10.21	24445	1834	PAR
10.21	742	110	PROCESSUS
10.08	90	28	TRADUCTEUR
10.00	814	116	BARTHES
9.88	41200	2929	DANS
9.88	329	62	MANUSCRIT
9.74	100	29	ALTÉRATION
9.64	17232	211213	DE
9.53	97	28	BECKETT
9.47	3333	328	DISCOURS
9.41	4757	437	NON
9.32	9035	748	S
9.18	114	30	
REFORMULATIO			
9.15	1682	189	MISE
9.12	168	38	UNITÉS
8.86	1869	202	RAPPORT
8.85	113	29	EXTRA
8.80	220	44	INDICES

Ecart	corpus	extrait	mot
8.79	154	35	DOCUMENT
8.68	239	46	DOCUMENTS
8.67	22345	1638	CE
8.61	6162	527	AUTRE
8.51	319	55	RÉFÉRENT
8.48	229	44	COUVERTURE
8.45	553	80	PRODUCTION
8.36	156	34	CODES
8.29	10638	836	QU
8.29	592	83	
INTERPRÉTATI			
8.25	3159	298	OEUVRE
8.17	1485	163	TITRE
8.10	613	84	PRODUIT
7.95	6697	553	ENTRE
7.93	1999	203	QUESTION
7.93	122	28	LITTÉRARITÉ
7.82	124	28	OPÉRATIONS
7.61	560	76	EFFETS
7.61	181	35	
FICTIONNELLE			
7.55	414	61	SOURCE
7.54	483	68	
LINGUISTIQUE			
7.52	724	91	STATUT
7.47	47747	3231	EST
7.45	16690	1222	COMME
7.37	179	34	PRAGMATIQUE
7.32	525	71	LECTEURS
7.29	43025	2921	UNE
7.26	190	35	
LINGUISTIQUE			
7.23	1431	150	JEU
7.15	423	60	NOTES
7.13	1345	142	TRAVAIL
7.08	67098	4431	À
7.06	723	88	CADRE
7.04	239	40	TROMPE
7.02	6340	510	C
6.96	224	38	
ARGUMENTATIO			
6.95	393	56	SÉMANTIQUE
6.87	3709	320	ÉCRITURE
6.86	850	98	NIVEAU
6.75	146	28	SÉMIOTIQUE
6.74	953	106	MODÈLE
6.72	1327	137	ENSEMBLE
6.71	287	44	GÉNÉRIQUE
6.70	116724	7495	LA
6.70	715	85	NARRATION
6.67	156	29	PRÉ
6.65	216	36	RÉGIME
6.64	38027	2572	QUI
6.61	1589	157	RÉEL
6.61	349	50	INSTANCE
6.53	176	31	NARRATIVES
6.52	3091	270	EFFET
6.48	258	40	
FONCTIONNEME			
6.46	178	31	
RÉFÉRENTIELL			

Graphique 4 : Environnement thématique du mot pôle TEXTE

À part les « corrélats » qui tiennent à la syntaxe (c'est le cas ici pour l'article masculin singulier *le* que la forme *texte* privilégie non à l'exclusion des autres articles *du*, *un*), à la phraséologie, aux expressions toutes faites, nous trouvons les véritables liens sémantiques,

le partage de sèmes communs par quoi on peut définir un thème. Le thème de la forme *Texte* dans notre corpus peut ainsi être défini à partir de l'usage que les études littéraires elles-mêmes en font et non plus par une norme externe. En haut de liste se trouvent des mots que l'on associe souvent avec texte : *lecteur*, *lecture*, *fiction*, *réception*, *traduction*, *référence*. Les co-occurrents sélectionnés sont eux-mêmes de très grande généralité, souvent très proches de *texte* comme *lecteur*, *fiction*, *réception*. C'est-à-dire que *texte* se prête à des mentions nombreuses dans des contextes très variés, difficiles à identifier lexicométriquement, voire peut-être même représentatif de l'ensemble du lexique du corpus. De cette hétérogénéité lexicale *TEXTE* tire les traits de généralité et d'abstraction sémantique qui le rendent, inversement, compatible avec tout prédicat. Ainsi le *TEXTE* comme élément de fond constitue une isotopie, produite par la récurrence de traits génériques, c'est-à-dire son isotopie générique dominante. Il assure ainsi aux différents co-occurrents une forte cohérence. « Le fond sémantique a une fonction d'unification d'une émanation du tout, dans la mesure où il définit la « matière » ou le « sujet » du texte », (2003 : 105), nous explique F. Rastier. Un fond sémantique ne se réduit pas à une isotopie, car il consiste en un faisceau d'isotopies, d'où la présence des isotopies spécifiques suivantes, qui restent complémentaires malgré leur spécificité :

1. Isotopie **narratologique :** « narrataire », « narration », « paratexte », « intertexte », « couverture », « titre », « littérarité », « Genette », « récit », « figure », « perspective », « contexte », « voix » ;
2. Isotopie **linguistique** : « référence », « sens », « discours », « linguistique », « pragmatique », « sémiotique », « énoncé », « énonciation », « signification », « structure », « signes », « valeur », « langage », « corpus », « langues » ;
3. Isotopie **littéraire :** « lecteur », « lecture », « intertextualité », « auteur », « lire », « Barthes », « manuscrit », « Métaphore », « genre », « cohérence », « romanesque », « narratif », « rhétorique », « esthétique », « description », « contenu », « citation », « lecture », « critique », « livre », « poétique », « traduction », « commentaire », « allégorie, « production » ;
4. Isotopie du domaine de **l'édition** : « manuscrit », « rapport », « livre », « document », « couverture », « titre », « version », « édition » ;
5. Isotopie « **philosophique** » : « Expérience », « acte », « imagination », « imaginaire », « monde », « réflexion », « vérité », « nation », « univers », « perception », « origine », « « illusion », « réalité » ;
6. Isotopie « **psychanalytique** » : « complexe », « plaisir », « réel », « sujet ».

Mais encore une fois aucune isotopie métalinguistique qui développe des énoncés définitoires de la forme « texte ». Et ce magnétisme autour du mot pôle « texte » est presque le même dans un autre mot pôle, qui n'est autre que le pluriel de « texte ». En d'autres termes, la structure thématique de la forme *Texte* a beaucoup de points communs avec celle de *Textes*. Cela montre encore une fois que le nombre n'a pas une grande influence sur la formation du lexique. Ce n'est pas donc un magnétisme provisoire qu'un autre aimant pourrait abolir. Cette indifférence s'explique encore une fois à notre avis par l'absence d'une théorie du texte dans les études littéraires, et par l'abondance des renvois intertextuels, comme en témoignent le contexte de ces deux formes cité tout le long de cet exposé.

Ecart	Corpus	Extrait	Mot
226.53	1969	1984	TEXTES
23.76	41114	2458	DES
20.75	48606	2693	LES
17.08	629	105	LITTÉRAIRES
16.59	6012	470	TEXTE
16.37	164	46	AUTOBIOGRAPH
15.51	2059	211	LECTURE
15.30	172322	7672	DE
15.11	481	81	ÉCRITS
14.66	6674	478	CES
14.00	116	33	FICTIONNELS
12.60	14515	833	OU
12.45	124	31	PUBLIÉS
11.91	267	47	CORPUS
11.38	656	80	AUTEURS
11.06	3076	232	FICTION
10.87	41200	1965	DANS
10.62	2281	182	LITTÉRATURE
10.61	3271	238	AUTRES
10.60	814	88	BARTHES
9.91	440	56	RECUEIL
9.81	97	22	BECKETT
9.68	990	95	CERTAINS
9.66	149	28	ROMANESQUES
9.49	72	18	ANTÉRIEURS
9.39	624	68	ÉCRIVAINS
9.28	2072	158	LITTÉRAIRE
9.03	7753	442	SONT
8.96	463	54	TRADUCTION
8.50	422	49	GENRES
8.49	1730	132	SOUVENT
8.32	379	45	FICTIONNEL
8.30	3646	232	ILS
8.11	155	25	NARRATIFS
8.04	188	28	MANUSCRITS
7.98	4653	278	LEUR
7.86	94	18	TRADUCTIONS
7.80	432	47	CRITIQUES
7.70	6118	344	DEUX
7.62	189	27	ESSAIS
7.61	1675	122	CRITIQUE
7.54	682	63	OEUVRES
7.53	128	21	THÉORIQUES
7.51	3144	198	AUTEUR
7.50	654	61	ÉTUDE
7.46	395	43	TELS
7.43	28537	1309	D'
7.22	287	34	GÉNÉRIQUE
7.20	3130	194	LECTEUR
7.08	401	42	PLUPART
7.04	207	27	METTENT
6.89	132	20	ALLUSIONS
6.88	288	33	LANGUES
6.78	6396	343	AUX
6.77	553	51	PRODUCTION
6.77	157	22	INTERTEXTUAL
6.71	407	41	ÉDITION
6.69	3159	190	OEUVRE
6.69	231	28	GENETTE
6.68	1006	78	ROMANS
6.50	3140	187	ONT
6.44	1192	87	GENRE
6.39	269	30	RÉFÉRENTIEL
6.38	1038	78	AUTOBIOGRAPH
6.37	1311	93	CEUX
6.30	1065	79	ANALYSE
6.27	525	47	LECTEURS
6.27	234	27	POÉTIQUES
6.23	137	19	COLETTE
6.21	451	42	EXEMPLES
6.16	394	38	DIVERS
6.15	1624	108	TROIS
6.15	1604	107	ÉCRIVAIN
6.05	1485	100	TITRE
6.05	229	26	ORAL
6.03	724	58	STATUT
6.01	179	22	PRÉSENTENT
5.95	800	62	AUTOBIOGRAPH
5.92	1350	92	ANNÉES
5.89	379	36	RÉCEPTION
5.77	695	55	NOTAMMENT
5.74	403	37	OUVRAGE
5.66	536	45	ESSAI
5.64	153	19	ÉCRITURES
5.60	247	26	MODES
5.55	947	68	RÉCITS
5.52	448	39	PEREC
5.45	420	37	QUESTIONS
5.45	297	29	ÉTUDES
5.44	28163	1230	D
5.44	172	20	CONTEMPORAIN
5.35	287	28	PUBLICATION
5.30	1999	120	QUESTION
5.27	321	30	LESQUELS
5.25	152	18	PROPOSER
5.23	386	34	NOMBREUX
5.23	179	20	ÉDITIONS
5.16	237	24	MODÈLES
5.15	769	56	PEUVENT
5.14	238	24	ORALE
5.09	1327	85	ENSEMBLE
5.08	813	58	PARTICULIER
5.00	604	46	POÈMES
5.00	466	38	E
4.99	622	47	DIFFÉRENTS
4.99	501	40	CONCEPTION
4.99	305	28	FRAGMENTS
4.98	1049	70	FRANÇAIS
4.96	161	18	RESTENT
4.91	263	25	UNS
4.91	190	20	LINGUISTIQUE
4.89	73768	3020	ET
4.89	279	26	TRAVAUX
4.89	191	20	PARU
4.86	1121	73	SIÈCLE
4.83	222	22	ROLAND
4.80	936	63	LIRE
4.69	594	44	NOUVELLES
4.68	613	45	PERSPECTIVE
4.67	855	58	DAUDET
4.62	37129	1562	QUE
4.62	231	22	TÉMOIGNAGE
4.61	1869	108	RAPPORT
4.56	280	25	DERNIERS
4.54	7850	371	SES
4.50	283	25	THÈSE
4.48	15169	674	SUR
4.48	575	42	TRADITION
4.48	300	26	THÉORIQUE
4.48	192	19	PRÉSENTATION
4.40	546	40	DIFFÉRENTES
4.35	889	58	ÉLÉMENTS
4.31	1032	65	UNIVERS
4.30	343	28	AUTOFICTION
4.28	344	28	APPROCHE
4.28	262	23	FRONTIÈRES
4.27	413	32	SOUVENIRS
4.25	186	18	ÉNONCIATIF
4.20	188	18	XIXE
4.14	642	44	PUBLIC
4.12	495	36	VERSION
4.10	255	22	TITRES
4.07	445	33	CONTES
4.00	560	39	EFFETS
3.96	453	33	TOUTEFOIS
3.91	663	44	SÉRIE
3.90	16690	722	COMME
3.90	232	20	FICTIONS
3.82	950	58	BEAUCOUP
3.81	850	53	NIVEAU
3.80	24445	1030	PAR
3.80	502	35	IMPORTANCE
3.80	339	26	PERMETTENT
3.79	674	44	DIMENSION
3.79	254	21	OUVRAGES
3.78	503	35	PREMIERS
3.73	343	26	TEXTUELLE
3.72	44966	1837	EN
3.69	242	20	ARTICLES
3.68	1752	95	LEURS
3.65	278	22	ÉTABLIR
3.65	261	21	NOUVEAUX
3.58	6697	307	ENTRE
3.58	389	28	CONCERNE
3.56	9371	417	NOUS
3.54	581	38	GRANDS
3.54	373	27	COMMENTAIRE
3.54	216	18	ÉCART
3.52	268	21	CATÉGORIES
3.49	825	50	PRATIQUE
3.48	1897	100	CAS
3.47	1181	67	ÉGALEMENT
3.47	306	23	TANTÔT
3.46	2940	146	MOINS
3.46	707	44	PROBLÈME
3.42	592	38	INTERPRÉTATI
3.42	592	38	CERTAIN
3.39	1715	91	QUELQUES
3.37	2299	117	AVANT
3.36	3709	178	ÉCRITURE
3.36	242	19	TEXTUEL
3.33	1244	69	PARTIR
3.32	315	23	IMPORTANT
3.31	1118	63	SAVOIR
3.29	2361	119	SOIT
3.29	263	20	HISTORIQUES
3.28	246	19	DONNENT
3.27	507	33	DOMAINE
3.26	469	31	PROSE
3.25	230	18	VASTE
3.24	1807	94	TANT
3.23	266	20	INVENTION
3.23	231	18	PRÉCIS
3.21	672	41	FRANÇAISE
3.19	10180	443	AVEC
3.19	1528	81	AUTANT
3.19	341	24	PHRASES
3.17	43025	1739	UNE
3.17	904	52	PARFOIS
3.15	698	42	ÉNONCIATION
3.14	865	50	ORIGINE
3.10	67098	2670	À
3.08	444	29	RELATIONS
3.03	588	36	PARMI
3.01	549	34	VOIRE
3.00	631	38	NOMBRE
3.00	392	26	POÈTES
3.87	2090	112	EXEMPLE
3.84	203	18	RÉÉCRITURE
3.83	390	29	PUBLIÉ

Graphique 5 : Environnement thématique du mot pôle TEXTES

Contrairement à des articles linguistiques, où le nombre permet la distinction entre le concept et le nom, le sens et les sens, la signification et les significations, le discours et les discours, l'interprétations et les interprétations, dans notre corpus la distinction s'estompe.

Tout est bon pour l'étude du texte et des textes, linguistique, philosophie, psychanalyse, mais un métalangage propre aux études littéraires en tant que tel n'existe pas.

En haut de liste se trouvent des mots que l'on associe souvent avec *texte* : lecteur, lecture, fiction, traduction, référentiel, réception, référence, narrataire, paratexte, commentaire, intertexte, auteur, implicite, roman, univers, genre, représentation, cohérence, etc. On s'y attendait, or des mots comme « notion », « concept », « discours » n'ont pas de place. C'est qui explique, encore une fois, la sous-représentation des expressions comme « notion de texte », deux occurrences, mais sans aucune activité définitoire,

> « En premier lieu il nous semble qu'une pièce de théâtre a un texte alors qu'un film n'en a pas, si du moins nous prenons **la notion de texte** au sens strict : tissu verbal, continu, contexture serrée formant un ensemble qui se suffit ».
>
> « **La notion de texte** utilisée par commodité pour désigner ce qui est détaché au moyen d'espaces, de tirets ou ce qui est isolé sur une page, c'est-à-dire ce qui se présente comme unité arrêtée, recouvre une réalité beaucoup plus diffuse et moins circonscrite à l'instar des éléments d'un tableau ou de tableaux que l'on pourrait embrasser du regard en faisant jouer, alterner les simultanéités sans concevoir une ordination qui octroierait un sens ».

Or pour l'expression « concept de texte » ou « le texte comme concept » le corpus n'enregistre aucune occurrence. C'est le cas aussi des lemmes des verbes *être*, *définir*, *entendre*. La seule explication derrière ces statistiques est la suivante : la forme « texte » est tout simplement un élément de fond (par sa fréquence) qui n'est pas élaboré comme concept par les études littéraires. Contrairement à d'autres formes, comme « réception », « focalisation », « sujet », rythme », « fiction », « référent », « représentation » pour lesquelles il y a un travail de théorisation. Or pour texte, la forme la plus suremployée, les études littéraires estiment que la théorie n'est pas de leur ressort. Car, contrairement à d'autres notions qui ont souvent été définies par telle ou telle école, et dont les définitions sont rappelées dans les articles, la notion de texte, elle, est employée sous le mode de l'évidence. On faisait comme si tout le monde savait de quel sens elle était porteuse. Or depuis Hjelmslev, la notion ne cesse d'intriguer. Dans toutes les théories qui se veulent textuelles, de la sémiotique à la littérature, en passant par la linguistique textuelle et la narratologie, les théoriciens du texte ne théorisent pas de la même façon. Les objectifs sont différents, l'heuristique est différente, du coup le sens du concept de « texte » n'est pas un bloc monolithique, mais un espace variée où se meuvent différentes écoles et plusieurs esprits.

7. CONCLUSION

La description sémantique met en évidence le caractère rigoureux d'une approche textuelle des formes. Les formes se construisent dans des textes, dans des contextes, que seuls les outils lexicométriques nous permettent d'observer. La description des formes n'a de valeur que si l'on peut rendre compte de leur évolution et leur description morphosémantique. En d'autres termes, l'interprétation des formes textuelles ne peut se faire sans l'examen des formes expressives auxquelles les formes sont liées. Le but premier était de montrer que les formes textuelles peuvent être étudiées comme des thèmes. En effet, une même forme se rencontre dans des genres différents, mais rien n'assure qu'elle se rapporte aux mêmes thèmes, car elle n'a pas la même molécule sémique, ni les mêmes lexicalisations. Une

forme comme « texte » devient le thème principal de notre corpus, elle revêt ainsi le statut de parangon, mais seulement comme élément de fond, et non comme concept. Pas d'énoncés définitoires autour de la forme texte. Ce n'est pas parce qu'une forme réalise des occurrences importantes dans un corpus qu'elle acquiert le statut de concept. Certes dans tous les articles de notre corpus, la forme « texte » est employée, mais à aucun moment elle n'accède, malgré sa dominance, au statut de concept. Du coup le texte dans les études littéraires est un texte sans textualité. Car la spécificité d'une forme n'est pas seulement caractérisée par les mots qui reviennent fréquemment dans son entourage mais également par la rareté ou l'absence de certains mots liés à l'activité définitoire. Du coup les textes analysés l'ont été dans une perspective qui définit les structures textuelles en relation directe avec l'auteur, le lecteur et le monde. La signification du texte reste donc conçue comme une suite d'énoncés, sans aucune considération pour les structures transphrastiques, notamment celle du genre ou des discours. En d'autres termes, la conception du texte dans les études littéraires se borne au mot et à la phrase et ne considère jamais le texte comme un palier de description. La textualité reste à définir et à spécifier, en dépassant les objectifs restreints du lecteur, de la fiction, de la réception, de la référence et de la narration. Car le texte comme le souligne F. Rastier, n'est pas un ensemble de phrases, « chaque texte procède d'un genre, et chaque genre est relatif à un discours » (1996 : 16).

8. REFERENCES

Brunet, E. 2002. *Hyperbase Version 5,2.* http://ancilla.unice.fr/~brunet/pub/hyperbase.html.

Loiseau, S. & Poudat, C. & Ablali, D. 2006. 'Exploration contrastive de trois corpus de sciences humaines' in *JADT 2006*, Presses Universitaires de Franche-Comté.

Biber, D. 1993. 'Using register-diversified corpora for general language studies' in *Computational Linguistics*, 19(2), p. 243-258.

Rastier, F. 1996. 'La sémantique des textes. Concepts et applications', *Hermes*, n°16, p. 15-37.

Rastier F. 2001. *Arts et Sciences du texte*. Paris : Presses Universitaires de France.

Rastier, F. 2003. 'Formes sémantiques et textualité' in *Unité(s) du texte*, *Cahiers du Crisco*, n°12, p. 99-114.

LA SÉMANTIQUE LEXICALE ET LES GENRES : ANALYSE SYSTÉMATIQUE D'UN CORPUS QUÉBECOIS

Margareta Kastberg Sjöblom
ILF-CNRS
Bases, Corpus et Langage (UMR 6039), Nice

1. INTRODUCTION

Le présent exposé propose d'étudier les variations et les oppositions génériques dans une base de données multigénérique québécoise, en s'appuyant sur un corpus informatisé et lemmatisé, et en exploitant les techniques quantitatives. Ce corpus présente une riche variété de textes et de corpus oraux et se décline en différents genres.

Le recours à des outils de traitement informatiques ou mathématiques particuliers permet de dégager les modèles de fonctionnement (lexical, grammatical, syntaxique, phraséologique, pragmatique) des discours et des textes pour une meilleure compréhension des genres discursifs et de la textualité. L'analyse des corpus en situation montre en effet que le lexique, la morphosyntaxe, la structure et le rythme du récit varient avec les genres. L'opposition entre les différentes typologies est toujours présente et souvent même prépondérante dans les différentes analyses statistiques, notamment lorsque l'on s'intéresse comme ici à la sémantique et aux structures thématiques.

Afin de compléter les possibilités de recherche déjà offertes par la mise en réseau du corpus par le Secrétariat à la politique linguistique du Québec, et étant donné la riche variation typologique des textes et la relative homogénéité de taille des différents sous-corpus, ce corpus nous a semblé particulièrement intéressant à soumettre à un traitement lexicométrique. L'exploration statistique du logiciel donne la possibilité d'analyses diverses non seulement traditionnelles comme celles sur la richesse lexicale, l'accroissement lexical, la distance lexicale, la corrélation chronologique etc., et il donne aussi l'accès à plusieurs fonctions thématiques. Le programme permet en effet d'analyser les spécificités lexicales aussi bien d'un point de vue endogène qu'exogène, il permet d'extraire les corrélats thématiques et recense aussi tous les termes situés dans l'environnement immédiat d'un mot donné. L'analyse des corrélats sémantiques et thématiques révèle des caractéristiques de chaque typologie présente dans ce corpus et l'analyse factorielle montre que les mêmes orientations de textes se retrouvent aussi bien aux niveaux lexical et syntaxique qu'au niveau thématique.

2. LE CORPUS

Parmi les grands projets de recherche sur la langue française au Canada, le secrétariat à la politique linguistique du Québec, grâce au Programme de développement, met en réseau et propose l'exploitation de corpus lexicaux québécois. Ce réseau est constitué de quinze corpus provenant de cinq universités québécoises : l'Université de Laval, l'Université de

Montréal, l'Université du Québec à Montréal, l'Université du Québec à Rimouski et l'Université de Sherbrooke.

Le corpus vise à établir, à partir de l'observation, le français standard du Québec, comme l'exprime son promoteur Pierre Martel du Conseil de la langue Française au Québec :

> *« La variété de français socialement valorisée que la majorité des Québécois francophones tendent à utiliser dans les situations de communication formelle. »*

Le but ultime de ce grand travail est la réalisation d'un dictionnaire du français au Québec, qui serait au Québec ce qu'est le *TLF* en France. Il s'agit de constituer un corpus représentatif du français au Canada (pas à partir des dictionnaires, surtout ceux venant de France) et d'y observer les usages.

Mais le projet québécois diffère dans la composition du corpus par rapport au *TLF*. En effet, le *TLF* a été fondé principalement sur des textes littéraires, même s'il contient quelques textes techniques (20 % au total). Ce rapport est inversé dans le corpus québécois dans lequel la part littéraire ne dépasse pas 15 %. La cible n'est pas le français soutenu, ni le français populaire, mais une norme « standard » dont la littérature, mais aussi les journaux, l'administration, les essais techniques, portent témoignage.

La base que nous nous proposons d'exploiter ici est un des sous-corpus de *La Banque de Données Textuelles de Sherbrooke*. Ce corpus contient environ deux millions d'occurrences (61 843 formes) tirées de 1 054 textes différents, et elle constitue donc un sous-ensemble de la *Banque de Données Textuelles de Sherbrooke* (BDTS), qui contient plus de 16 millions d'occurrences à l'heure actuelle.

Ce corpus est composé de huit sous-ensembles d'environ 250 000 mots chacun et traités selon une norme commune, ce qui rend leurs données comparables.

Les sous-corpus sont représentatifs de divers domaines, types de discours et niveaux de langue et la répartition est la suivante : 10 % de textes techniques, 14 % de textes scientifiques, 11 % de textes sociopolitiques, 13 % de textes administratifs, 16 % articles de journaux, 13 % de textes littéraires, 13 % de textes environnementaux, et 10 % d'oral provenant d'enquêtes sociolinguistiques.

Les textes techniques sont composés d'extraits de rapports, guides, manuels de formation, normes, procédures, etc. dans les domaines de l'aluminium, de l'environnement, des mines, des domaines de pâtes et de papier, de la télécommunication, du transport, de l'hydro-électricité et de l'informatique. Les textes scientifiques sont constitués de cent textes extraits de mémoires, de thèses, d'articles scientifiques et de rapports de recherche dans les domaines de la biologie, de la chimie, de la physique, du génie chimique et mécanique. Quant aux textes sociopolitiques, ils sont constitués de cent textes sélectionnés à partir des 583 mémoires présentés à la Commission, notamment des mémoires ou extraits de mémoires de la Commission Bélanger-Campeau (la consultation populaire sur l'avenir du Québec). Le sous-corpus administratif est composé de 54 textes issus de débats de l'Assemblée nationale, de conventions collectives, de textes juridiques, de textes du Bureau d'audiences publiques sur l'environnement (BAPE) et de rapports du vérificateur général du Québec. L'unité journalistique est constituée à partir de 52 textes tirés de *L'Actualité*, 55 textes provenant du *Devoir* (quotidien économie, Québec, Ottawa), 17 textes tirés *d'Interface*, 63 textes tirés de *La Presse (*quotidien de Montréal), 57 textes provenant de *Québec Science* et 75 textes tirés du *Soleil quotidien* (information Québec). Le sous-corpus littéraire est issu de 25 textes ; des romans, des chansons, des essais, des textes de poésie et des pièces de théâtre. Nous y trouvons notamment le roman *L'ange*

exterminé de Gérald Godin (1990), les recueils de nouvelles *Le cassé et autres nouvelles* de Jacques Renaud (1982), *Contes sur la pointe des pieds* de Gilles Vigneault (1960). Parmi les pièces de théâtre nous y trouvons *La Dalle-des-Morts* de Felix Antoine Savard (1965), et enfin l'essai critique littéraire *Ecrire de la fiction au Québec* de Noël Audet (1990). Les textes environnementaux sont représentés par 107 textes divers du ministère de l'Environnement du Québec, des journaux, des périodiques et des magazines spécialisés dans ce domaine. Finalement, le sous-corpus oral est composé des enquêtes sociolinguistiques effectuées dans la région des Bois-Francs (centre du Québec, à mi-chemin entre Québec et Montréal).

Ce corpus témoigne non seulement d'une grande qualité technique mais aussi d'une extrême richesse par la large variation typologique des textes ainsi que la relative homogénéité de taille des différents sous-corpus.

Afin de compléter les possibilités de recherche déjà offertes par la mise en réseau de ce corpus, il a semblé particulièrement intéressant de le soumettre à un traitement lexicométrique, permettant notamment de cibler certaines caractéristiques sémantiques.

3. MÉTHODES ET TECHNIQUES

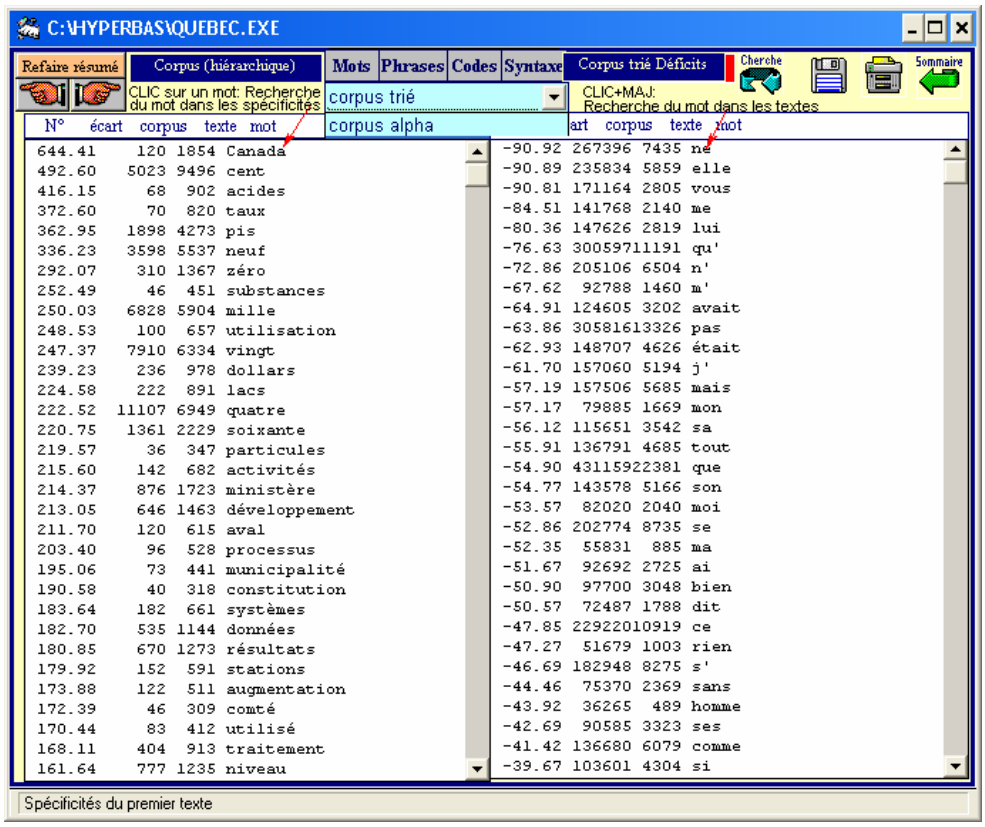
Le logiciel *Hyperbase* est un logiciel d'exploitation documentaire et de traitement quantitatif des grands corpus textuels qui autorise un ensemble de traitements sur des corpus de textes prédéfinis et numérisés. Il permet ainsi le traitement statistique et automatique des données, qui peut servir de plateforme pour diverses études sur un corpus défini ainsi que pour la comparaison avec les données du corpus *Frantext*.

Hyperbase permet non seulement l'exploitation documentaire et l'obtention des contextes et des concordances, mais de plus la distribution d'un mot peut être étudiée dans l'ensemble des textes qui composent le corpus de travail et visualisée grâce aux applications graphiques. L'exploration statistique du logiciel donne la possibilité d'analyses diverses, non seulement traditionnelles comme celles sur la richesse lexicale, l'accroissement lexical, la distance lexicale, la corrélation chronologique etc., mais plusieurs fonctions thématiques sont également disponibles. Le logiciel *Hyperbase* permet d'analyser les spécificités lexicales d'un point de vue aussi bien endogène qu'exogène. Il permet d'extraire les corrélats thématiques et recense aussi tous les termes situés dans l'environnement immédiat d'un mot donné, d'extraire le réseau isotopique. *Hyperbase* est aujourd'hui en mesure de traiter les sorties du lemmatiseur-étiqueteur *Cordial* pour la langue française et de *TreeTagger* pour l'allemand, l'anglais et l'italien. Il ne permet donc plus seulement de traiter les « mots », mais aussi les lemmes, les codes grammaticaux, les enchaînements syntaxiques ou les corrélats sémantiques.

4. ANALYSES LEXICO-SÉMANTIQUES

La thématique des différents récits peut être exploitée, avec le recours au logiciel logométrique *Hyperbase*, de façon très précise, en s'intéressant aux spécificités des différents récits, à leur évolution et à leurs contextes.

L'analyse des spécificités est une démarche classique, que le logiciel accomplit en s'appuyant sur *Frantext*, et plus précisément sur le corpus du XXe siècle. Les mots qui se trouvent en tête de liste de la partie gauche de la figure ci-dessous mettent en relief la spécificité et la réalité quotidienne de la vie canadienne par rapport aux mots « franco-français » du corpus *Frantext*.



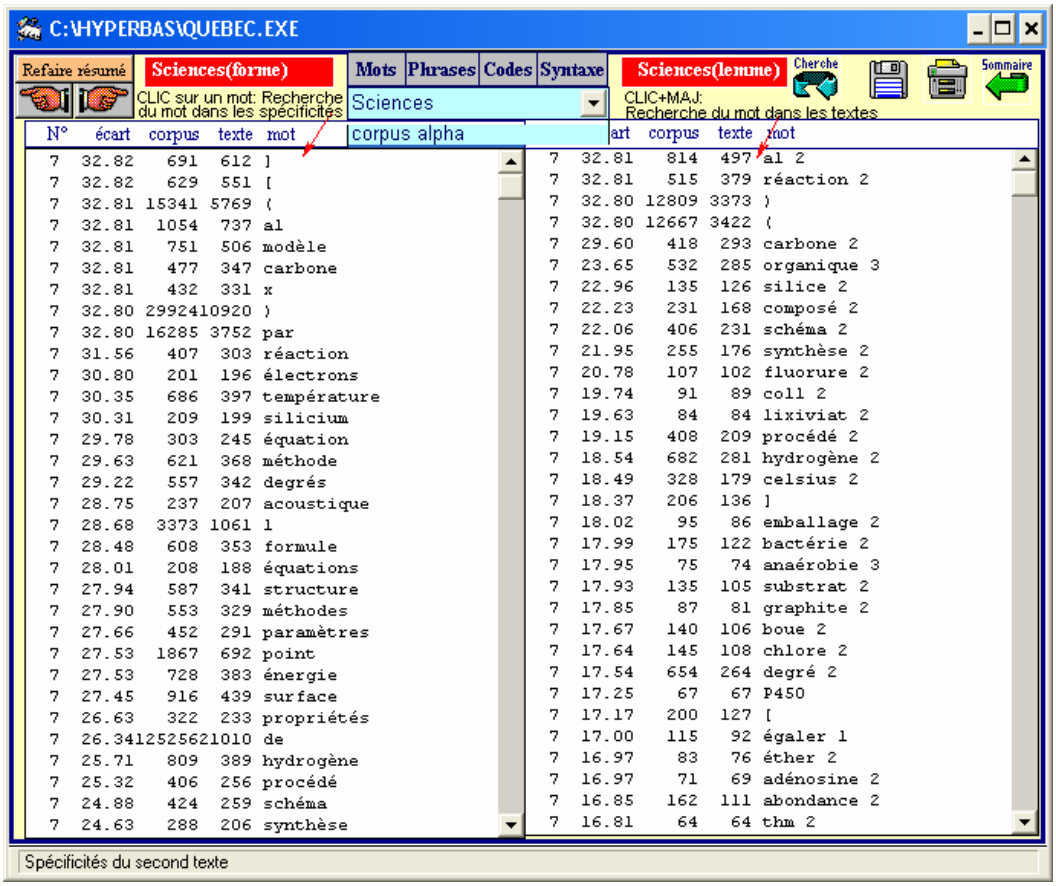
Graphique 1 : Spécificités lexicales de la ***BDTS***

Nous trouvons également des mots relatifs à la vie parlementaire et administrative, moins employés dans le corpus français qui puise une large partie de ses textes dans la littérature.

On peut constater que cette liste ne contient pas de verbe, ni même d'adjectif ou d'adverbe. Il n'y a aucune articulation du discours : la phrase semble constituée de substantifs juxtaposés. Il s'agit ici évidemment d'un effet de genre textuel, reflété par la grande présence d'un discours intellectuel, faisant appel au substantif.

À droite du graphique, nous trouvons les spécificités négatives, c'est-à-dire les mots statistiquement sous-employés dans le corpus *Québec*. Ici, nous pouvons constater encore le même effet du genre avec un déficit important de pronoms personnels, bien plus représentés dans les textes littéraires que dans ce corpus multigénérique.

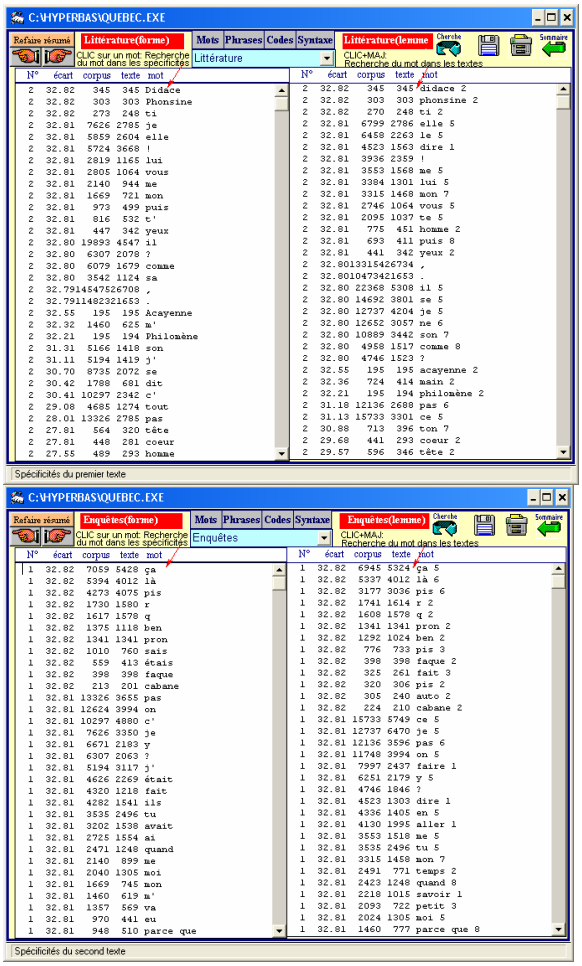
Le logiciel permet également l'observation du vocabulaire spécifique de chacun des sous-corpus, c'est-à-dire une comparaison endogène. Cette spécificité est déterminée par le calcul de l'écart réduit pour chaque forme dans chaque partie du corpus. Les textes sont comparés, les uns après les autres, avec le corpus dans son ensemble. Ces comparaisons internes se justifient facilement, puisque le corpus est expressément conçu pour mettre en valeur les différences qui opposent les textes dans ce même ensemble. S'il est homogène – comme ici – le calcul relèvera toujours des écarts intéressants que l'œil humain ne pourrait saisir.



Graphique 2 : Spécificités lexicales du sous-corpus scientifique

Les résultats sont très nets, ces mots reflètent parfaitement le genre du discours et nous donnent le profil caractéristique de chaque typologie. Les spécificités lexicales du genre scientifique nous affiche des lexèmes tels que : *modèle*, *carbone*, *réaction*, *électrons*, *température*, *silicium*, *équation* et *méthode*. Le corpus administratif est caractérisé par des mots comme *article*, *employé*, *règlement*, *employeur*, *syndicat*, *leader*, *indemnité*, *motion* et *alinéa*, ce qui n'est pas très étonnant. Dans le corpus « Environnement », il est souvent question de l'Europe avec les pays européens en tête de liste, tels *France*, *Paris*, *Italie*, *Espagne*, *Allemagne*, *Angleterre*, qui apparaissent avant des termes comme *vins*, *peinture* et *festival*. Le corpus journalistique est caractérisé par des mots qui appartiennent à ce monde : *illustration*, *photo*, *article* etc., ainsi que des mots qui reflètent la vie courante canadienne comme *dollars*, *Montréal* et *Soleil*. Rien d'étonnant à trouver des mots tels que *carbone*, *réactions*, *électrons*, *température* et *silicium* en tête de liste des spécificités du sous-corpus scientifique. En revanche, le parallélisme du corpus littéraire et du corpus oral l'est peut-être davantage.

Le discours des deux sous-corpus fait appel aux mêmes pronoms des deux premières personnes, aux démonstratifs *ça* et *ce*, aux verbes d'état ou courants comme *être*, *faire*, *dire* et *penser*. Il semble en effet que le discours littéraire contemporain présent dans notre corpus soit proche de l'oral, bien que les spécificités du langage familier québécois, comme *pis*, *ben*, *pron* et *faque*, soient moins représentées, comme le montrent les deux histogrammes suivants :



Graphique 3 : Spécificités lexicales du corpus littéraire et du corpus oral

Le logiciel permet aussi l'observation de tous les mots présents dans l'entourage immédiat d'un mot choisi pour pôle et la comparaison de la fréquence de ces corrélats dans ce sous-corpus, constitué par l'entourage immédiat, avec celle du corpus entier, pour ainsi extraire l'univers lexical ou bien le réseau isotopique qui entoure un mot.

Il s'agit d'un calcul de spécificité particulier, puisqu'on ne recherche plus une relation entre un mot et un texte, mais une relation privilégiée entre les mots eux-mêmes – ce que mesure aussi le calcul de corrélation, quand deux séries sont juxtaposées.

C:\HYPERBAS\QUEBEC.EXE

N°	écart	corpus	texte	mot
290.69	1854	1818		CANADA
41.47	393	125		ANGLAIS
40.42	6419	593		QUÉBEC
31.06	373	93		UNIS
31.04	28364	1329		DU
28.83	631	116		RESTE
26.23	582	102		ÉTATS
24.86	15	14		GENDARMERIE
24.39	68	30		OCÉANS
24.03	79	32		PÊCHES
23.14	117	38		RADIO
22.28	16	13		RÉVISÉES
21.29	14864	676		AU
20.53	144	38		BANQUE
20.41	355	62		PROVINCES
20.18	1071	116		PAYS
19.10	40	18		ROYALE
18.13	122	31		STATISTIQUE
17.33	1269	114		ENVIRONNEMEN
17.33	10	8		MCQ
17.30	64	21		SUPRÊME
17.24	51597	1622		LE
17.09	84	24		BELL
15.49	27	12		MT
15.30	181	33		MEECH
15.22	60	18		NÉGOCIER
15.19	317	45		SOCIAL
13.73	51	15		SOUVERAINIST
13.63	459	51		CANADIENNE
13.36	1025	82		POLITIQUE
12.95	460	49		NATIONALE
12.89	357	42		ÉMISSIONS

N°	écart	corpus	texte	mot
4.13	638	28		1985
5.00	115	10		1ER
6.33	148	14		2000
12.14	313	37		ACCORD
3.55	193	11		ACTUELLE
3.22	326	15		ACTUELLEMENT
3.40	593	24		AFFAIRES
4.14	101	8		AFFIRME
6.59	231	19		AGRICULTURE
4.66	885	38		AILLEURS
5.20	167	13		AMÉRICAIN
3.20	185	10		AMÉRICAINS
4.23	233	14		AMÉRIQUE
41.47	393	125		ANGLAIS
8.03	70	11		ANGLETERRE
8.26	67	11		ANGLOPHONE
8.32	89	13		ANGLOPHONES
3.11	1086	37		ANNÉE
6.84	118	13		ANNUELLE
9.02	383	33		ASSOCIATION
8.28	102	14		ASSOCIÉ
21.29	14864	676		AU
5.69	67	8		AUTOCHTONES
7.85	6747	231		AVEC
7.28	606	38		AVENIR
20.53	144	38		BANQUE
3.12	525	21		BAS
17.09	84	24		BELL
3.45	3048	90		BIEN
9.86	28	8		BILINGUE
6.38	98	11		BRITANNIQUE
3.08	164	9		BRUT

Spécificités du second texte

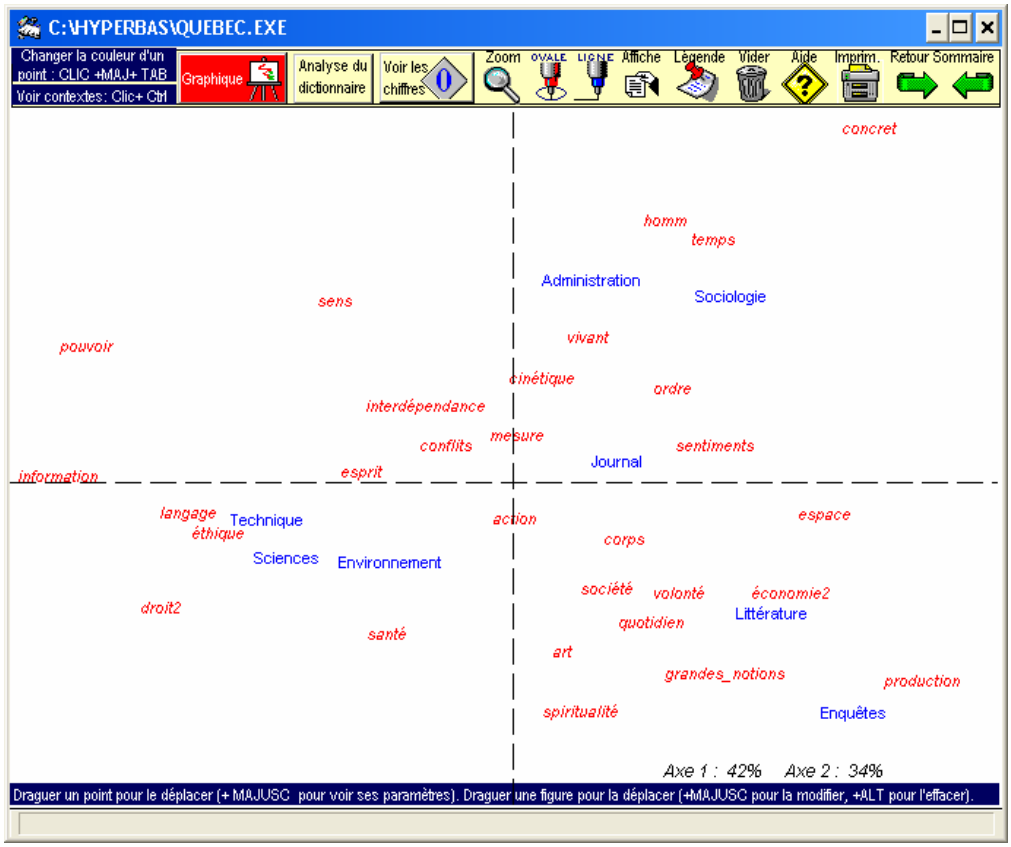
Graphique 5 : Univers thématique du mot Canada

Le clivage générique est donc toujours présent et les résultats montrent des caractéristiques bien particulières, spécifiques des différents genres.

De la même façon, l'analyse des corrélats sémantiques et thématiques révèle aussi les caractéristiques de chaque typologie présente dans ce corpus et l'analyse factorielle montre que les mêmes orientations des textes se retrouvent aussi bien aux niveaux lexical et syntaxique qu'au niveau thématique.

Le traitement sémantique proposé par *Cordial* vise à rendre compte des idées, des sentiments, des actions, c'est-à-dire des thèmes exprimés dans un texte. Le programme *Cordial* fait en effet appel à un thésaurus de référence, où sont cataloguées les disciplines, les concepts et les connaissances. Il convient ici de souligner une certaine fragilité de l'analyse due à l'homographie et même à la polysémie, et sans doute ces étiquettes s'appliquent-elles mal à certains corpus.

Pourtant, malgré ces faiblesses et les incertitudes du codage sémantiques, les résultats auxquels celui-ci conduit ne sont pas sans intérêt. On retrouve au niveau sémantique le même clivage générique et la même « aimantation » de textes qu'on observe aux niveaux lexical et syntaxique.



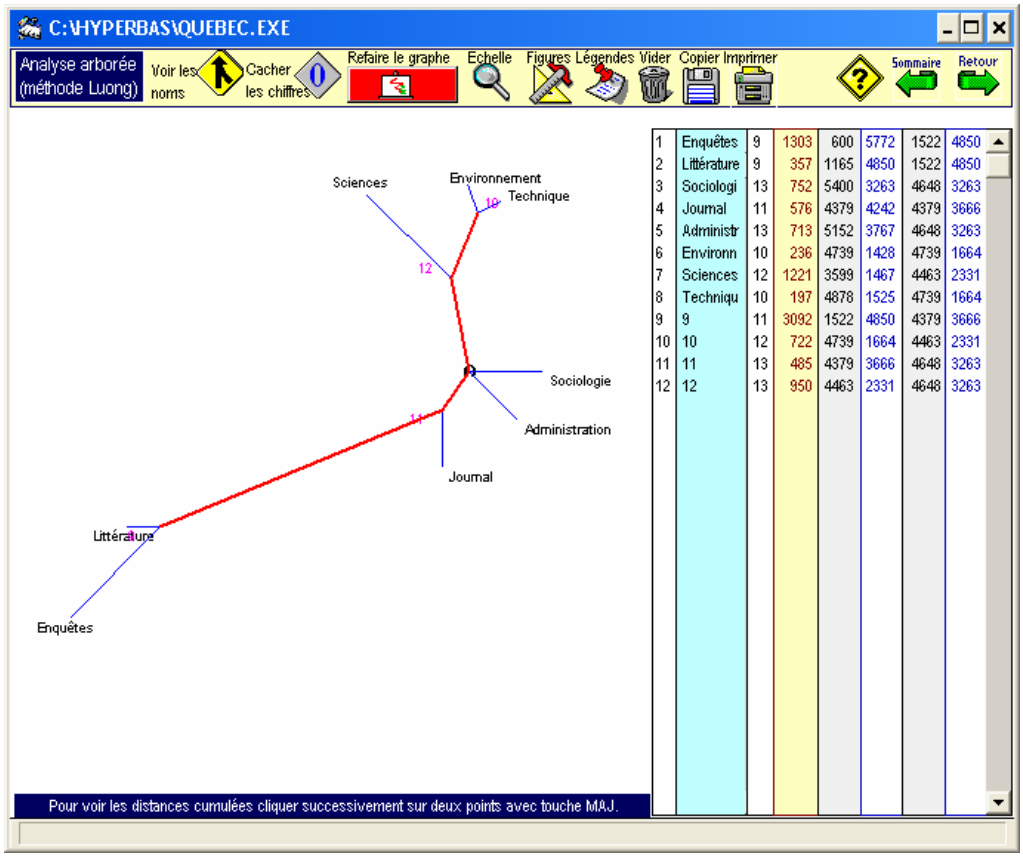
Graphique 6 : Analyse factorielle des thèmes

Une autre façon d'aborder genres et thèmes est l'étude de la distance lexicale ; il s'agit de considérer le vocabulaire intégral de chacun des textes du corpus et de rassembler ceux qui partagent des thèmes semblables.

Il est possible de calculer la distance lexicale entre chaque sous-corpus, en prenant en compte tous les mots rencontrés et leur répartition, partagée ou exclusive dans les textes. D'une manière générale, la distance absolue entre deux textes A et B sera l'union de ces deux textes moins leur intersection, soit : $(A \cup B) - (A \cap B)$.

En réalité il existe plusieurs façons de faire ce calcul : on peut tenir compte ou non de la fréquence ; si on travaille sur V, on utilise la formule Jaccard, et on ne s'intéresse qu'à la présence ou à l'absence prenant appui sur les lemmes ou sur les graphies.

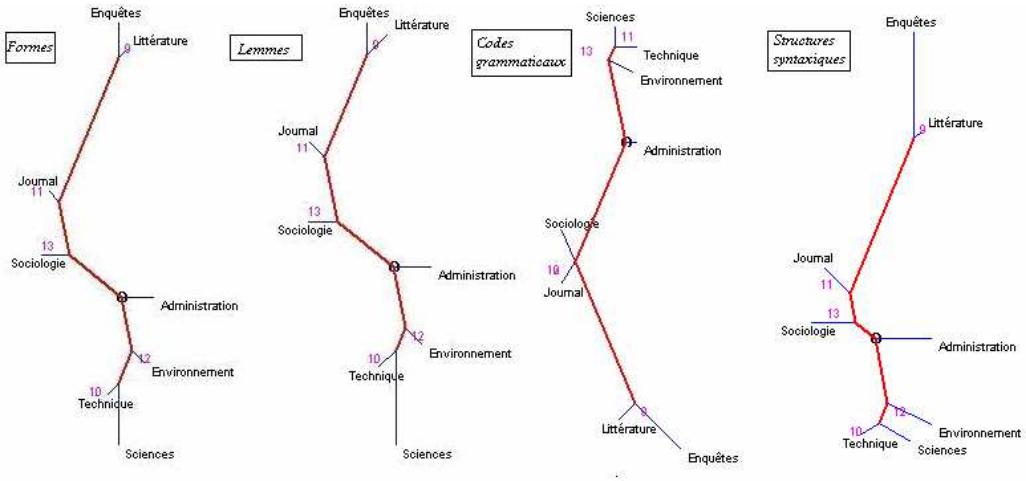
La figure ci-dessous dessine la carte typologique qui résume l'ensemble de ces mesures de proximité et d'éloignement, en prenant considération de la fréquence le calcul N (formule Labbé) :



1	Enquêtes	9	1303	600	5772	1522	4850
2	Littérature	9	357	1165	4850	1522	4850
3	Sociologi	13	752	5400	3263	4648	3263
4	Journal	11	576	4379	4242	4379	3666
5	Administr	13	713	5152	3767	4648	3263
6	Environn	10	236	4739	1428	4739	1664
7	Sciences	12	1221	3599	1467	4463	2331
8	Techniqu	10	197	4878	1525	4739	1664
9	9	11	3092	1522	4850	4379	3666
10	10	12	722	4739	1664	4463	2331
11	11	13	485	4379	3666	4648	3263
12	12	13	950	4463	2331	4648	3263

Graphique 8 : Analyse factorielle de la distance lexicale

Les analyses arborées ci-dessous montrent les résultats du calcul non seulement sur le lexique, mais aussi sur les codes grammaticaux et sur les structures syntaxiques. On voit que les résultats sont sensiblement les mêmes, mettant en avant les mêmes oppositions génériques.



Graphique 7 : Analyse arborée de la distance intertextuelle, présence/absence V

On observe ici que l'on obtient la même image, alors que les objets considérés sont étrangers les uns aux autres : formes, lemmes, codes grammaticaux ou enchaînements syntaxiques. Il est vrai qu'entre les formes et les graphies il existe une part commune, tout

comme pour les codes et les structures syntaxiques, ces dernières étant des combinaisons des codes. Mais quel lien existe-il entre les graphies et les structures ? On peut constater que les sous-corpus s'orientent différemment à tous niveaux, lexical, grammatical ou syntaxique, selon le genre.

En effet, les variables que l'on croit indépendantes sont liées par des accords ou par une commune soumission à une très forte influence, celle du genre ou de la typologie, prépondérante dans toute analyse quantitative. Chaque genre littéraire a en fait son anatomie, sa physiologie et son fonctionnement, et cela transparaît très clairement dans ce corpus québécois.

5. REFERENCES

Adam J.-M. 2005. Les textes types et prototype : Récit, description, argumentation, explication et dialogue, Paris : Armand Colin (coll. : « Fac. Linguistique ».

Brunet E. 2000. "Peut-on mesurer la distance entre deux textes ?", in Rastier F. (éd.) Corpus littéraires – Recueil et numérisation, analyses assistées, didactique, Paris 20-21 octobre 2000.

Kastberg Sjöblom M. 2002. L'écriture de J.M.G. Le Clézio, une approche lexicométrique, Thèse de doctorat, Université de Nice-Sophia Antipolis.

Kastberg Sjöblom M.& Brunet E. 2000. "La thématique. Essai de repérage automatique dans l'œuvre d'un écrivain", in Rajman M. & Chappelier J.-C. (éds.), JADT 2000, 5èmes Journées internationales d'Analyse statistique des Données Textuelles, Ecole polytechnique fédérale de Lausanne, p. 457-465.

Malrieu D. & Rastier F. 2002. "Genres et variations morphosyntaxiques", in Angel Martin Municio (éd.), 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, p. 61-84.

Williams G**. (ed.) 2005. *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes**.

ORGANISATION INFORMATIONNELLE DU DISCOURS ORAL NARRATIF EN F.L.E. : ÉTUDES DE QUELQUES CAS EN MILIEU HOMOGLOTTE

Hanitra MAURY
Université de Bretagne Sud et Doctorante
CRINI (EA 1192), Axe Didactique des langues et cultures étrangères, Université de Nantes

1. LA COMPETENCE DISCURSIVE ET SA COMPLEXITE

Trouver des critères qualitatifs dans le but d'évaluer l'expression orale reste une éternelle problématique pour l'enseignant, car cette habilité langagière met en jeu d'autres paramètres tels que la personnalité, l'âge, la motivation de l'apprenant. De plus, la complexité et l'hétérogénéité du discours s'expliquent par le fait que c'est un produit d'interactions de plusieurs informations lexicales, syntaxiques, textuelles et situationnelles[1].

À titre illustratif, prenons l'exemple du descripteur qualitatif de l'utilisation de la langue parlée proposée par le Conseil de l'Europe2. Une nécessité de clarification s'impose au niveau de la compétence discursive3 dans la partie cohérence dont les niveaux4 de compétences à acquérir sont décrits de la manière suivante :

L'apprenant

> **A1** « Peut relier des mots ou groupes de mots avec des connecteurs très élémentaires tels que 'et' ou 'alors'. »
>
> **A2** « Peut relier des groupes de mots avec des connecteurs simples tels 'et', 'mais' et 'parce que'. »
>
> **B1** « Peut relier une série d'éléments courts, simples et distincts en une suite linéaire de points qui s'enchaînent. »
>
> **B2** « Peut utiliser un nombre limité d'articulateurs pour lier ses phrases en un discours clair et cohérent, bien qu'il puisse y avoir quelques 'sauts' dans une longue intervention. »

En apparence, le descripteur semble renvoyer à une simple analyse qualitative à la surface de ces articulateurs. Pourtant, leur interprétation relève également d'une autre dimension, celle de la macrostructure, liée directement aux systèmes hiérarchique et référentiel du discours, en l'occurrence à l'identification de la notion citée dans la littérature (Brown & Yule, 1983 ; Givon, 1983) comme étant les « fils conducteurs », les « leitmotiv », les « à propos de » et que nous appellerons dans ces travaux les *topiques*.

Berrendonner (1983) a avancé l'hypothèse que ces informations n'étaient pas souvent explicitées dans le discours oral, bien qu'elles aient leur source dans l'environnement cognitif immédiat ou dans les connaissances encyclopédiques des interlocuteurs. L'enchaînement de ces informations est assuré grâce, entre autres, aux connecteurs[5] et aux anaphores. Figurant parmi les marques linguistiques des topiques, ces derniers jouent un rôle prépondérant dans leur identification. Une grande partie de cette communication sera

ainsi consacrée aux expressions référentielles, et plus particulièrement aux pronoms dans l'introduction ou le maintien du topique, c'est-à-dire l'information immédiate se trouvant dans la mémoire discursive des interlocuteurs et leur servant de point d'ancrage (Grobet, 2002).

2. DES LOGICIELS D'ANALYSE AUTOMATIQUE AU SERVICE DE L'ENSEIGNANT ?

La question n'est pas de remettre en cause le descripteur proposé par le Conseil de l'Europe, mais de réfléchir sur son application possible avec l'introduction d'un cadre de référence et l'apport d'instruments d'analyse dont la crédibilité a été déjà évoquée dans la littérature (Roulet, 1999 ; Grobet, 2002).

Se situant même au cœur du processus de formation, l'évaluation est une phase indispensable pour la progression de l'apprenant et pour soutenir sa motivation. Cependant, l'évaluation de la production orale et sa correction restent très difficiles à accomplir pour diverses raisons dont une non négligeable, le manque de temps. À l'heure où les nouvelles technologies prennent de plus en plus de place dans l'enseignement / apprentissage des langues, il convient de se demander s'il n'est pas judicieux de mettre au service de la didactique des logiciels de traitement automatique des langues naturelles.

> « […] l'analyse des propriétés et des régularités de l'organisation d'un texte constitue un préalable à toute interprétation […] »[6].

En quoi l'apport de l'outil informatique pourrait-il être d'utilité pertinente à l'enseignant, voire à l'apprenant ? Nous tenterons de répondre à cette question en utilisant TROPES, un logiciel d'analyse lexicologique, sémantique et discursive. Etant donné que le discours est composé de systèmes, appelés également modules, cette étude devrait être réalisée en plusieurs étapes : linguistique, textuel et situationnel. Les modules lexical, syntaxique, hiérarchique et référentiel seront traités successivement dans le but de trouver leur interrelation et d'identifier les paramètres de certains processus assurant la cohérence et la cohésion des discours. Cette démarche est fondée sur le fait qu'il est difficile, parfois, pour des apprenants de langues étrangères, de connaître à fond les fils conducteurs, les thèmes, les leitmotivs et les personnages dans un flux discursif. Ce constat pourrait provenir également des difficultés liées à l'acquisition de certains systèmes assurant la cohérence et la cohésion du discours, tels que la pronominalisation (cohérence référentielle), la linéarisation et l'acquisition de certains connecteurs (cohérence relationnelle)[7].

2.1. Description du logiciel TROPES

Le choix du logiciel TROPES[8] a été particulièrement motivé de par sa conception et afin de répondre à des analyses cognitivo-discursives de discours écrits ou oraux et à leur interprétation. C'est un logiciel qui s'appuie sur une logique d'Intelligence Artificielle (IA) et qui utilise des dictionnaires et des réseaux sémantiques de classification pour analyser qualitativement et quantitativement des données textuelles.

À travers l'analyse de la référence, il détecte le contexte, les thèmes et les principaux acteurs du discours. Il identifie son style et sa mise en scène et classifie les verbes, adjectifs, adverbes, pronoms personnels et conjonctions en métacatégories sémantiques. Cette analyse qualitative est complétée par une étude chronologique. Ainsi, le discours est divisé en grands épisodes, ce qui assure le suivi de l'évolution *des propos du discours* et

permet de pouvoir en dégager les systèmes sous-jacents. La vérification de toute hypothèse d'interprétation peut être entreprise grâce à de nombreux graphes hypertextes auto-organisés. Cette étude vise à démontrer l'intérêt de tel logiciel pour le formateur, en tant qu'outil supplémentaire d'évaluation en production orale (ou écrite) lui offrant alors la possibilité de choisir des moyens d'accompagnement adaptés. En outre, il peut servir d'outil d'auto-évaluation à l'apprenant, l'aidant à découvrir les acquis, les erreurs ainsi que certaines stratégies d'apprentissage, tout en favorisant son investissement et le développement de son autonomisation.

Cela étant, ce logiciel n'est pas un outil d'évaluation optimal, mais doit être un instrument qui viendrait parfaire les enregistrements ou les grilles d'évaluation afin de rendre l'évaluation de l'oral la moins subjective possible. Cela n'effacerait pas non plus l'intuition de l'évaluateur liée à ses expériences et à ses connaissances. Par ailleurs, une réflexion sera engagée sur les limites de TROPES dans ces investigations.

2.2. Corpus

Le corpus étudié est celui d'apprenants FLE dans un centre universitaire en France[9] et issus de différentes nationalités : chinoise, brésilienne, polonaise, argentine, jordanienne, thaïlandaise, taiwanaise. Nous analyserons un court fragment de cet enregistrement pour introduire le concept de complexité du discours et pour examiner de plus près la forme d'organisation informationnelle.

Par ailleurs, étant donné qu'une étude décontextualisée serait, pour des raisons méthodologiques, inadéquate, le thème choisi est celui du « script du restaurant ». Les étudiants ont eu dans cette procédure à décrire en premier lieu six images de scènes se déroulant dans un restaurant et ensuite, après un débat issu d'une différence d'interprétation d'une séquence[10], six autres images illustrant la suite. Le choix du thème semble anodin ; pourtant il permettra, d'une part, de vérifier des hypothèses émises à partir des résultats d'analyse constitués par TROPES, telles que l'existence d'un système conceptuel ou référentiel de base sous-jacent au niveau du script et au niveau des catégories (Jagot, 2002) et, d'autre part, ce système référentiel nous aidera à comprendre les procédures possibles pour maintenir dans une continuité linéaire l'information concernant un référent identifié et activé, *l'entité topicale* (Givon, 1983 ; Grobet, 2002).

2.3. Exemple de trascriptions [11]

1Ka : Je m'appelle Ka / je suis Brésilienne\\

euh / ce que je vois / dans cette BD / sont / deux messieurs \\ un serveur et un monsieur qui / qui est assis pour manger / dans un supposé restaurant // euh / il est bien servi = il mange // ensuite / euh .. le serveur le présente l'addition / et / il va jusqu'à la caisse / pour payer cette addition \ et là il cherche / il fouille dans sa poche / pour chercher l'argent \ et ensuite \ on voit qu'il / il enlève les poches / il montre qu'il n'a pas d'argent \\ et / et on voit le / le / le serveur qui le / le montre l'addition / il doit lui demander euh = comment il va faire / pour payer cette addition \\ voilà / et ça / ça suit comme ça l'histoire \\

3. ANALYSE DU DISCOURS

3.1. Système linguistique[12]

3.1.1. ANALYSE LEXICALE Les recherches concernant l'analyse des structures syntaxiques du français oral ont démontré que le marquage des topiques se manifeste le plus souvent à l'aide soit des éléments lexicaux tels que les anaphores il, elle, le etc.., soit des démonstratifs parmi lesquels l'élément pertinent ça, soit des syntagmes nominaux indéfinis et définis (Grobet, 2002). Retrouvera t-on les mêmes marquages ou tout au plus des indices de marquage à la surface dans le corpus étudié et comment interpréter le référent auquel se rapporte ces expressions ?

TROPES qualifie le discours étudié plutôt d'argumentatif. Il a été pris en charge par le narrateur, ce qui justifie l'usage à 34 % du pronom *je*. Cependant, ces mêmes résultats classent *il* avec un taux de 54 % comme étant l'expression référentielle la plus usitée par les étudiants. Cela signifie t-il que le propos du discours se concentre autour de cet élément ?

En examinant de plus près le graphe étoilé[13] qui affiche clairement les relations entre ce pronom et ces référents, nous pouvons remarquer que son fondement ne repose pas sur une relation sémantique :

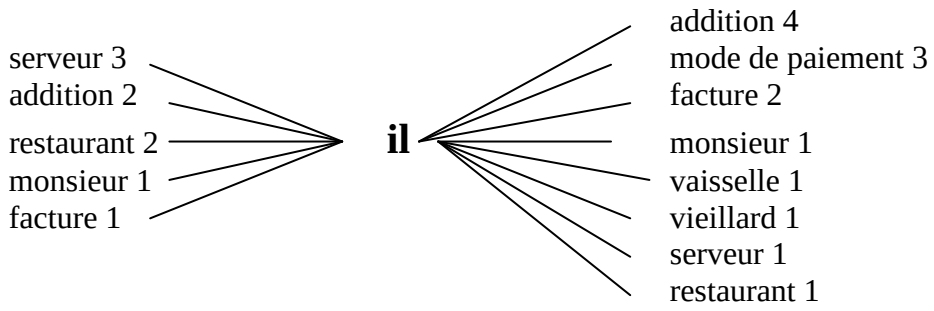


Tableau 1 : Le graphe étoilé du pronom il

En effet, d'après ce graphe, les références affichées à gauche précèdent *il* et celles de droite le suivent. La portée des observations de TROPES se limite à une analyse de l'environnement de *il*, ce qui pourrait rendre le concept de référence ambigu. Une analyse textuelle, classique, du pronom personnel clitique *il* considèrerait toutes les références affichées à gauche comme étant des antécédents accessibles et identifiés, c'est-à-dire[14]

> « […] un référent déjà connu par l'interlocuteur, c'est-à-dire un référent « présent » ou déjà manifeste dans la mémoire immédiate […] ».

De plus, cette analyse s'appuyant sur la notion de voisinage contextuel ne prend pas non plus en compte la valeur sémantique de *il*, masculin, singulier, qui est un des facteurs justifiant la distribution des pronoms, et de ce fait exclut ici obligatoirement les références telles que *addition* et *facture*.

Le fait que TROPES présente les références sans déterminant soulève également une autre difficulté dans notre démarche pour l'identification des topiques[15]

> « L'identifiabilité concerne […] le statut d'un référent dans la connaissance des interlocuteurs : un référent est dit identifiable, si sa connaissance est supposée partagée par l'interlocuteur et s'il est verbalisé comme tel […]. La catégorie de

l'identifiabilité permet de rendre compte du contraste entre les expressions référentielles indéfinies (syntagme nominal ou pronom indéfini) et les expressions définies (pronoms, syntagmes nominaux définis, démonstratifs ou accompagnés de l'adjectif possessif). »

1Ka : euh / ce que je vois / dans cette BD / sont deux messieurs \\ **un serveur** et **un monsieur qui / qui** est assis pour manger / dans un supposé restaurant \\ euh / **il** est bien servi = **il** mange // euh / **il** est bien servi = **il** mange // ensuite / euh .. **le serveur le** présente l'addition / et / **il** va jusqu'à la caisse / pour payer cette addition […].

Dans cet extrait, Cet étudiant respecte cette notion d'identifiabilité en utilisant au début les syntagmes nominaux indéfinis *un serveur et un monsieur* pour introduire l'élément « non connu », repris tout d'abord par *qui*, puis par le pronom *il*. Un peu plus loin, la reprise de *un serveur* par *le serveur*, un syntagme nominal défini a pour objectif de réactiver un référent déjà identifié.

La différence fondamentale entre un pronom non accentué comme *il* et un nom accentué comme *le serveur* repose sur leur état d'activation ou d'accessibilité au moment de l'énonciation. Les pronoms marquent généralement une plus forte accessibilité que les syntagmes nominaux[16]. Les deux premiers *il* font référence à *qui* ayant pour antécédent *un monsieur*. L'utilisation du singulier peut être expliquée par la saillance de ce référent dont la source est à chercher plus dans une conception cognitive : le serveur ne peut pas être, dans ce contexte, assis ou bien servi. En revanche, le référent du troisième *il* peut être ambigu, car il pourrait être aussi bien *le serveur* ou le pronom *le*. Toutefois, on n'a pas d'autres informations sur l'état d'activation du syntagme *le serveur*. Une étude complémentaire fondée sur le principe de pertinence pourrait justifier toute autre interprétation significative. Cela nécessiterait alors une analyse de la syntaxe dans laquelle apparaissent ces éléments.

3.1.2. ANALYSE SYNTAXIQUE Il faut souligner encore une fois que le topique est un concept difficile à cerner, car il en existe moult définitions (Grobet, 2002 ; *Cahiers de praxématique* n°40). Nous emprunterons ici la définition présentée par Grobet (2002) : la notion de *topique* en tant qu'information accessible, identifiée et activée, ce qui signifie l'information immédiate se trouvant dans la mémoire discursive des interlocuteurs et leur servant de point d'ancrage.

Dans la structure canonique assertive *Sujet verbe complément*, la fonction de topique est attribuée généralement au sujet. En effet, l'information structurelle de la proposition étant neutre, des variations contextuelles peuvent donc mener à diverses interprétations. Le savoir contextuel et situationnel partagé rend le discours oral beaucoup plus économique que l'écrit, la difficulté réside dans le décodage et dans l'encodage de l'information. Le canal oral impose une linéarité stricte à l'écoute et à la production. En outre, beaucoup d'informations peuvent être implicites. Par conséquent, des structures prosodiques et syntaxiques typiques, à l'instar des clivées ou des segmentées, servent en français à marquer les topiques :

1 Ka : […] sont / deux messieurs \\ un serveur et un monsieur qui est assis pour manger dans un supposé restaurant […]

C'est une structure clivée où *ce* n'est pas explicité. Le français parlé a tendance à ne pas conjuguer ni accorder *c'est*[17], ce qui n'est pas le cas dans cet exemple.

1 Na : […] mais / le serveur il demande el pourquoi […]

La segmentée à gauche sert à activer le référent *le serveur*, ce qui aide à lever toute ambiguïté dans l'interprétation du topique.

À propos du logiciel TROPES, il occulte l'importance de la structure syntaxique dans son analyse discursive, d'autant plus que les critères déterminant les propositions remarquables s'appuient sur des analyses éclatées des catégories et des références[18].

3.2. Système hiérarchique

Après cette analyse du système linguistique, passons à une autre étape, celle du système hiérarchique en liaison directe avec l'organisation relationnelle, c'est-à-dire avec le rôle des connecteurs, utilisés par TROPES pour l'interprétation typologique du texte. Nous insistons sur le fait qu'il pourrait avoir plusieurs interprétations possibles et que grâce à l'utilisation de cet outil heuristique et à ce couplage entre informations hiérarchiques et relationnelles, il est possible de suivre la progression informationnelle.

La structure hiérarchique du discours analysé est constituée d'interventions et d'actes principaux et secondaires :

I	Ap euh <u>ce</u> que je vois (pseudo-clivée) *Préparation* As dans cette BD *cadre (toile de fond)*
I	Ap sont deux messieurs un serveur et un monsieur <u>qui</u> qui est assis pour manger *topicalisation* As dans un supposé restaurant *cadre*
I	Ap euh <u>il</u> est bien servi <u>il</u> mange ……..

Tableau 2 : La structure hiérarchique de l'exemple 1Ka

Chaque acte principal introduit l'objet du discours qui correspond à l'information servant de point d'ancrage dans la mémoire discursive c'est-à-dire le topique. Verbalisée, cette information pourrait être explicitée, par exemple, de la manière suivante :

À propos de la personne qui est assise, c'est un monsieur …

Il est possible d'avancer l'hypothèse que la source du topique se trouve aussi bien dans l'avant-texte , grâce aux images, mais aussi dans l'environnement cognitif des étudiants présents au moment de l'enregistrement.

Quant à l'acte, *il est bien servi il mange*, le topique est identique à celui de l'acte précédent, ce qui démontre la présence d'une progression linéaire. Il est marqué par le pronom anaphorique *il*. S'appuyant sur l'analyse thématique présentée par TROPES, l'étude d'une autre dimension, notamment celle du système référentiel est, par la suite, indispensable afin de démontrer l'existence d'un enchaînement à topique constant.

3.3. Système référentiel

3.3.1. LE SCRIPT TROPES définit la notion de *référence* comme étant le contexte, les grands thèmes, répondant à la question *sur quoi porte le texte ?*

Présentant *la restauration* comme thème principal du discours. il serait intéressant dans une approche cognitive, à travers les classifications ainsi que l'analyse des graphes étoilés ou des scénari des références, de détecter une structure conceptuelle sous-jacente au niveau du script du restaurant. Cette hypothèse repose sur des théories en psychologie cognitive (Abbott *et al.*, 1985) caractérisant le script comme partie intégrante de la mémoire sémantique et répondant à des principes d'organisations de *partie / tout* et *de niveau de base d'abstraction*.

Le terme « restauration », classé en premier dans l'univers des références par TROPES, implique des représentations afférentes, telles que « serveur », « client », « plat », « addition », auxquelles sont associées[19]

> « des actions de niveau hiérarchique élevé, baptisées 'titres de scène' […] ».

Cette conception est largement admise par les linguistes (Roulet, Grobet) qui partagent également l'idée d'activation des représentations associées à une structure référentielle, dite également conceptuelle.

> « Cette structure repose en partie sur des connaissances générales encyclopédiques, culturelles et sociales acquises par l'expérience antérieure du locuteur et que l'on suppose partagées à l'intérieur d'une même communauté culturelle ».[20]

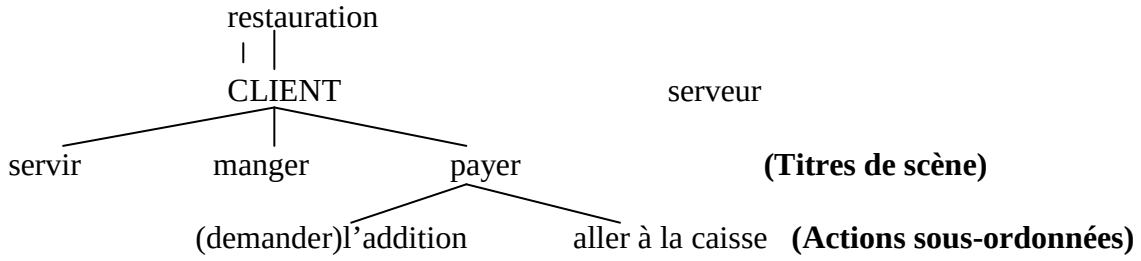


Tableau 3 : La structure conceptuelle de l'exemple 1Ka

3.3.2.L'ENTITE TOPICALE Brown & Yule (1983 : 137) définit l'entité topicale comme étant le personnage, l'objet ou l'idée principale d'un segment du discours qui est le plus souvent codée par le sujet ou par un élément en début de phrase. Cette fonction de sujet de l'entité topicale est partagée par Givon (1983 : 8) qui, en outre, la présente en tant que le *leitmotiv* du discours. À propos de notre exemple, il pourrait être *le monsieur*, le référent évoqué souvent par *il* dont le degré de saillance est discutable : *il* pourrait être également considéré dans ce cas comme point d'ancrage d'arrière-fond.

4. CONCLUSION

Bien que le logiciel TROPES présente des lacunes pour procéder à une analyse automatique du discours, il peut servir d'outil intéressant pour évaluer la compétence lexicale. Par ailleurs, l'élargissement de son usage à d'autres types de discours d'apprenants serait primordial. Comme on a pu le constater, l'analyse du discours implique différentes étapes d'interprétation, ce qui, dans une perspective didactique, changerait grandement non seulement la conception des critères d'évaluation mais également celle de

la progression dans l'enseignement / apprentissage des langues. La dimension interactionnelle n'a pas été évoquée, quoiqu'elle joue un rôle important dans la construction de systèmes de références. Mettre au service de l'enseignant un logiciel de reconnaissance vocale, qui transcrirait directement les données, améliorerait incontestablement sa gestion de travail sur l'oralité. Par conséquent, l'utilisation d'outils informatiques est une nécessité et leur développement lui donnera un coup de pouce.

5. NOTES

1. Nous nous référons à l'approche genevoise et à Roulet pour une description détaillée de l'organisation du discours. Par ailleurs, il est indispensable de préciser l'importance des systèmes syntaxique, hiérarchique et référentiel dans cette forme d'organisation.
2. Mis B., 2002, « L'enseignement des langues vivantes en Europe : le défi de la diversification », in *Les Cahiers du CIEP*, Didier, p. 32.
3. Inspiré du modèle SPEAKING de l'ethnologue de la communication D. H. Hymes, le Conseil de l'Europe présente la compétence de communication en trois composantes : linguistique, socioculturelle et discursive. Se définissant comme étant la charnière des composantes linguistique et discursive, la composante socioculturelle semble être le socle pour acquérir les autres.
4. Les niveaux C1 et C2 ne seront pas traités dans cette communication, car ils ne sont pas inclus dans le corpus étudié.
5. Les connecteurs servent avant tout à marquer la relation entre les actes qu'ils introduisent et les informations en mémoire discursive (Grobet, 2002).
6. Roulet E**.** (1999), *La description de l'organisation du discours*, Paris : Didier, p. 24.
7. Maury H**.** (2003), « Le conflit cognitif comme processus indispensable à l'acquisition d'une langue étrangère », in *Actes du colloque dynamiques du conflit* du CRELLIC, Lorient 20,21, 22 novembre 2002.
8. TROPES peut être téléchargé gratuitement sur le net. Malheureusement, certaines fonctions importantes, comme imprimer, copier et coller, ne sont pas disponibles dans cette version. Une explication détaillée est lisible sur le site http://www.acetic.fr..
9. Il s'agit d'étudiants chercheurs en Master 2 en Sciences de l'Education, option Didactique des Langues Etrangères, de l'université de Nantes.
10. Une longue discussion a été entamée autour de l'expression *pourboire* dont la source de la divergence serait plutôt d'ordre culturel.
11. Code de transcription utilisé :
(X), (XX), (XXX) syllabe ou suite de syllabes incompréhensibles
. ; .. ; … pauses plus ou moins longues
\ ; \\ ; / ; // ; = la direction des barres indique la direction de la mélodie
\ ; / indique un intonème continuatif
\\ ; / indique un intonème conclusif
(…) Commentaires ajoutés tels que (rire)
12. Même si le rôle de l'intonation est incontestable dans le marquage du topique (Morel, 1992a ; Grobet, 2002), il n'est pas toujours évident de le démontrer, car chaque langue a sa spécificité pour coder la structure informationnelle.
13. Selon les explications, ce graphe paraît le plus adéquat pour cette étude, car il affiche les relations entre références et catégories des mots.
14. Kleiber G. (1994a), *Anaphores et pronoms*, Louvain, Duculot, p. 25.
15. Grobet, A. (2002), *L'identification des topiques dans les dialogues*, Bruxelles, Duculot, p. 133.
16. *Ibidem*, p. 130.
17. Blanche-Benveniste C. (1997), *Approches de la langue parlée en français*, Paris, Ophrys, p. 98.
18. Les propositions remarquables résument les parties les plus caractéristiques du discours.
Il faut comprendre le concept référence ici comme étant le contexte. Grâce à une analyse sémantique, les mots les plus fréquents sont identifiés et classés par thèmes avec leurs équivalents. Par exemple, TROPES a trouvé onze mots équivalents à l'un des grands thèmes de ce discours « restauration ».
19. Jagot L. (2002), « Catégories et scripts », in Cordier F. et François J. (dir.), *Catégorisation et langage*, Paris : Lavoisier, Hermès Science, p. 150.
20. Grobet A**.** (2002), *op. cit.*, p. 306.

6. RÉFÉRENCES BIBLIOGRAPHIQUES

Abbott V., Back J. B., Smith E. E. (1985), *The Representation of scripts in memory*, Journal of Memory and Language, 24, p. 179-199.

Berrendonner A. (1983), 'Connecteurs pragmatiques et anaphores', in *Cahiers de linguistique française 5*, p. 215-246.

Blanche-Benveniste C. (1997), *Approches de la langue parlée en français*, Paris : Ophrys.

Brown G. & Yule G. (1983), *Discourse Analysis*, Cambridge : Cambridge University Press.

Castelotti V., Py B. (coord.) (2002), *La notion de compétence en langue*, Lyon : ENS Editions.

Cordier F., François J. (dir.) (2002), *Catégorisation et langage*, Paris : Hermès.

Delbecque N. (éd.) (2002), *Linguistique cognitive. Comprendre comment fonctionne le langage*, Bruxelles : de Boeck Duculot.

Gaonac'h D., Larigauderie P. (2000), *Mémoire et fonctionnement cognitif*, Paris : Armand Colin.

Givon T. (éd.) (1983), *Topic Continuity in Discourse. A Quantitative Cross-Language Study*, Amsterdam / Philadelphia : Benjamins.

Grobet A. (2002), *L'identification des topiques dans les dialogues*, Belgique : Editions Duculot.

Hymes D. (1972), 'On communicative competence', in Pride J.B. et Homes J. (éd.), *Sociolinguistics*, Penguin Modern Linguistics Reading.

Jackendoff R. (1997), *The Architecture of the Language Faculty*, Cambridge Mass : MIT Press.

Jagot L. (2002), 'Catégories et scripts', in Cordier F. et François J. (dir.), *Catégorisation et langage*, Paris : Lavoisier, Hermès Science, p. 145-163.

Kleiber G. (1994a), *Anaphores et pronoms*, Louvain : Duculot.

Maury H. (2003), 'Le conflit cognitif comme processus indispensable à l'acquisition d'une langue étrangère', in *Actes du colloque dynamiques du conflit* du CRELLIC, Lorient 20,21, 22 novembre 2002.

Mis B. (coord.) (2002), *L'enseignement des langues vivantes en Europe : le défi de la diversification*, Paris : Didier, p. 8-49.

Neveu F. (coord.) (2003), 'Linguistique du détachement', in *Cahiers de praxématique*, n°40, Université Paul Valérie, Montpellier 3.

Pekarek-Doehker S. (Nov. 2001), 'Dislocation à gauche et organisation interactionnelle', in *Marges Linguistiques n°2*, Saint- Chamas : M.L.M.S. éditeur, p. 177 -194.

Roulet E. (1999), *La description de l'organisation du discours. Du dialogue au texte*, Paris : Didier / Hatier.

Schneidecker C., Theissen A. (coord.) (2001), 'Topicalisation et partition', in *Cahiers de praxématique*, n°37, Université Paul Valérie, Montpellier 3.

Vion R. (2000), La communication verbale, Analyse des Interactions, Paris : Hachette.

CONSTITUTION ET EXPLOITATION PÉDAGOGIQUE DE RESSOURCES LINGUISTIQUES POUR UN ENSEIGNEMENT/ APPRENTISSAGE DU DISCOURS RAPPORTÉ

Fanny Rinck, Thomas Lebarbé
Laboratoire LIDILEM, EA 609
Université Stendhal – Grenoble 3
{Fanny.Rinck ; Thomas.Lebarbe}@u-grenoble3.fr

1. INTRODUCTION

Le travail présenté ici vise la constitution et l'exploitation de ressources linguistiques pour l'enseignement/apprentissage du discours rapporté. Il s'inscrit dans le cadre d'un projet réunissant linguistes, talistes[1] et didacticiens avec comme point de départ la question de savoir comment favoriser l'enseignement/apprentissage des phénomènes énonciatifs, et plus précisément des phénomènes de manifestation de la subjectivité du locuteur et de polyphonie, c'est-à-dire de multiplicité des voix et points de vue dans les textes. L'idée est qu'une approche en corpus de ces phénomènes doit favoriser leur enseignement/apprentissage et il s'agit donc de constituer des corpus annotés exploitables à l'aide d'outils issus du TAL dans des scénarii pédagogiques destinés à différents publics, et adaptés à la spécificité de chacun :

- un public francophone de collégiens et lycéens, voire, pour certains des phénomènes, d'étudiants du supérieur ;
- un public d'étudiants de français langue étrangère (FLE), de niveau intermédiaire à avancé.

2. MOTIVATIONS DU PROJET

Les motivations didactiques de ce projet reposent sur le constat de lacunes dans l'enseignement/apprentissage des phénomènes énonciatifs. Quel que soit le public concerné, celui-ci se limite en général au discours rapporté et l'aborde de manière normative (Boch & Grossmann, 2001), en référence au bon usage, à partir d'exemples fabriqués par le linguiste ou le pédagogue qui représentent des structures typiques, restreintes au cadre phrastique ; les activités mises en place s'inscrivent dans un paradigme grammatical (cf. les exercices de transposition discours direct/ discours indirect), sans prise en compte du rôle joué par ces phénomènes dans la compréhension et la production de textes.

Ces constats induisent des besoins aussi bien sur le plan linguistique qu'en termes d'outils et d'automatisation de l'approche. Sans rejeter les approches traditionnelles de ces phénomènes en contexte didactique, il importe en effet de recourir à des données langagières attestées ; l'utilisation de corpus doit ainsi permettre d'aborder les phénomènes

de façon fonctionnelle, en tenant compte de la diversité des usages, au plan de la structure syntaxique, comme au plan sémantique avec un travail sur la subjectivité du locuteur, son point de vue et la polyphonie des textes. La diversité interne du corpus offre une prise sur la manière dont ces phénomènes varient avec les genres (roman, article journalistique…) et les types de texte (argumentatif, descriptif…). Le corpus enrichi d'annotations syntaxiques et sémantiques permet par ailleurs la formulation de requêtes et la génération automatisée d'exercices. Il s'agit donc de développer des approches fondées sur le repérage, l'analyse des phénomènes aux plans syntaxique et sémantique et leur manipulation, et d'intégrer les exercices traditionnels à un cadre plus large, à même de favoriser une réflexion métalinguistique des apprenants dans une perspective d'aide à la compréhension et à la production de textes.

3. DISCOURS RAPPORTÉ

Nous nous centrerons ici sur un seul des phénomènes énonciatifs qui intéressent notre projet, le discours rapporté. Celui-ci est encore peu traité dans une perspective d'annotation[2], *a fortiori* à des fins d'exploitation didactique des corpus annotés. Pourtant, il est depuis longtemps décrit par la grammaire traditionnelle selon une tripartition morphosyntaxique en fonction du rapport entre le discours citant et le discours cité, et en fonction du marquage du discours cité :

1- discours direct (type X a dit : « … ») :
"En avant !" s'écria le reporter[3] ;

2- discours indirect (type X a dit que…) :
Moi, qui voulais en savoir plus long, je demandai au voiturier de monter à côté de lui[4] ;

3- discours indirect libre (*i.e.* sans marque signalant le discours cité) :
Iscariote hésite un instant, puis, prenant un Arabe par le pan de son burnous :
- Tu l'as vu, Achmed, tu l'as vu... tu étais là. Le chrétien m'a frappé... Tu seras témoin... bien... bien... tu seras témoin.

L'Arabe dégage son burnous et repousse le juif... ***Il ne sait rien, il n'a rien vu : juste au moment, il tournait la tête...***[5].

Cette tripartition reste cependant discutable (Authier, 1979, 1992-93 ; Maingueneau, 1994) et selon nombre d'auteurs, c'est la conception même du discours rapporté qui demande à être repensée. Les approches plus récentes engagent à envisager le discours rapporté en termes de discours représenté (Rabatel, 2003), dans le sens où le texte met en scène un autre acte de langage, qui n'a pas toujours eu effectivement lieu[6]. Il est aussi question de « discours autres », ou de « discours d'autrui » (Boch & Grossmann, 2001), l'intérêt étant d'intégrer d'autres phénomènes proches, de manifestation de polyphonie ou de dialogisme dans les textes (Bres & Vérine, 2002), comme les modalisations en discours second, du type *selon X, d'après X, dixit x, paraît-il, il paraît que*, ou encore le discours narrativisé, qui consiste non à rendre compte des propos eux-mêmes, mais à représenter leur contenu de manière condensée, comme dans *les poules parlent de passer la nuit !...*[7].

Ces approches plus récentes ont donc en commun d'insister sur l'extrême diversité des formes de représentation de ces discours autres, et montrent que le discours rapporté n'est qu'un cas parmi d'autres du phénomène de polyphonie. En ce qui concerne la seule description morphosyntaxique du discours rapporté, elles optent quant à elles en faveur

d'un continuum dans les modes de rapport entre discours citant et discours cité. L. Rosier (1999) notamment défend cette idée de continuum, en montrant qu'il existe des degrés d'actualisation du discours citant, et que l'on peut trouver des formes mixtes entre discours direct et indirect (par exemple le discours direct avec *que*) ou encore des formes de discours direct libre.

Tout en souscrivant à ces approches plus récentes des phénomènes de discours rapporté et discours autres, notre perspective d'annotation à visée d'exploitation didactique appelle cependant à recourir à des catégories stables, établies à partir d'indices formels, d'ordre syntaxique et énonciatif. Cela étant, l'objectif est de rattacher le discours rapporté à d'autres phénomènes exclus des approches traditionnelles. On s'intéresse en particulier à des phénomènes qui dépassent le cadre strict du discours rapporté, comme les modalisations en discours second et le discours narrativisé (cf. *supra*), les segments introducteurs de discours rapporté, qu'on aborde du double point de vue de leurs caractéristiques morphosyntaxiques et sémantiques, et enfin le phénomène de la mise entre guillemets, qui recouvre pour une part celui du discours rapporté puisque les guillemets peuvent signaler une citation, mais dépasse aussi ce cadre, en ouvrant comme les segments introducteurs vers la question de l'expression du point de vue de l'auteur et celle de la polyphonie :

> La «**victoire**» zambienne devrait relancer le débat sur le recours aux OGM[8].

À plus long terme, l'objectif est d'intégrer également le lexique porteur de subjectivité (Kerbrat-Orecchioni, 1990) : les termes affectifs et axiologiques, ainsi que les modalités épistémiques, signalant le degré d'adhésion du locuteur à ce qu'il dit, qu'on a pour l'heure annoté uniquement dans le cadre des segments introducteurs de discours rapporté :

Et tout en me servant, le vieux me disait à l'oreille **d'un air de gourmandise** :

> - Vous êtes bien heureux, vous, de pouvoir en manger !... C'est ma femme qui les a faites... Vous allez goûter quelque chose de bon[9].

4. CORPUS

L'objectif pédagogique est de faire observer, analyser et manipuler en contexte certains phénomènes linguistiques. La constitution d'un corpus est donc un préliminaire nécessaire à cette application mais est soumise à de nombreuses contraintes dues autant à l'objectif applicatif qu'au phénomène que l'on cherche à modéliser et à enseigner.

4.1. Contraintes et définition du corpus

La première contrainte est d'ordre pédagogique. En effet, il est nécessaire, pour une application pédagogique telle que nous la concevons, de proposer un volume de données textuelles suffisamment large pour créer des activités pédagogiques qui ne soient pas répétitives. L'enseignant doit pouvoir créer de nombreux exercices sur un même phénomène sans pour autant utiliser à chaque fois les mêmes exemples en contexte.

D'autre part, ces activités doivent présenter une certaine diversité des usages, liée aux styles d'auteur, aux genres et aux périodes. Ces différents écrits doivent, de plus, offrir une large représentativité des phénomènes énonciatifs afin d'offrir un large panel d'activité par phénomène. Enfin, le corpus doit être conçu de manière à rendre les activités pertinentes et intéressantes pour le public concerné.

Une telle combinatoire de contraintes d'ordre pédagogique génère donc un corpus de taille conséquente dont la constitution est coûteuse.

Il existe toutefois une contrepartie à l'ambition d'un tel corpus. Si la législation française autorise « les analyses et courtes citations justifiées par le caractère […] pédagogique » (Code de la Propriété Intellectuelle – CPI – Article L-122-5 3a), elle ne permet pas la mise à disposition d'œuvres (« reproduction » au sens du CPI) encore sous le couvert du droit patrimonial, y compris par le biais d'une interface de base de données. En conséquence, trois possibilités se présentent afin d'intégrer des œuvres récentes[10] :

1. négocier à titre payant certaines œuvres
2. négocier à titre gratuit certaines œuvres
3. rejeter les œuvres récentes afin de ne pas attirer les foudres des éditeurs.

Afin d'allier représentativité et respect de la législation en vigueur, nous avons dû constituer deux corpus, l'un étant un sous-ensemble du premier.

Le corpus à visée de mise à disposition pédagogique en ligne est limité à des textes pour lesquels nous avons obtenu les droits ou tombés dans le domaine public. En revanche, le corpus à visée de modélisation scientifique du phénomène constitue le corpus englobant. Son utilisation étant uniquement dédiée à des fins d'observation et de modélisation, nous ne sommes pas contraints par la législation, si ce n'est de ne pas le diffuser.

Le premier est composé d'ouvrages littéraires et d'essais antérieurs au 20e siècle, d'écrits scientifiques et d'écrits glanés sur Internet, notamment des « Blogs » après obtention de l'accord de leurs auteurs[11]. Le corpus de recherche complète le précédent avec des écrits du même ordre et des journaux que nous ne sommes pas libres de diffuser. En définitive, le corpus pédagogique représente approximativement 450.000 mots et 4.500 phénomènes de discours rapporté.

4.2. Annotation

Dans ce cadre didactique, la notion de corpus n'est concevable qu'à partir du moment où il est manipulé à l'aide d'outils informatisés de recherche au sein des textes et de mise à disposition aussi bien pour l'enseignant que pour l'apprenant. Une telle approche implique une structuration et une annotation du corpus préalables à tout traitement.

La structuration du corpus (structuration en sources mais aussi structuration intrinsèque des documents eux-mêmes) dont on ne traitera pas ici se conforme à la Dublin-Core Metadata Initiative (DCMI : http://dublincore.com).

Une annotation du discours rapporté, objet de cette communication, est nécessaire de manière à en identifier systématiquement les éléments constitutifs à des fins de description linguistique mais aussi à des fins pédagogiques. Cette annotation est effectuée sous le format XML afin de permettre une extraction et une visualisation des occurrences et de leurs contextes. Le langage de description XML permet en effet une description formelle des contenus (DTD) ainsi que l'utilisation de feuilles de styles (CSS) – pour la mise en forme – et de modèles de transformation (XSLT) – pour la manipulation et *a fortiori* l'extraction.

Il n'existe pas d'outil de traitement de langue suffisamment performant et fiable pour une annotation automatique des phénomènes de discours rapporté, bien que des travaux soient effectués dans cette direction (Mourad et Declés, 2002). Les marqueurs physiques permettant de les identifier ne sont pas assez réguliers et les marques syntactico-sémantiques n'ont pas encore été formalisées, si tant est qu'elles puissent l'être (Rosier, 1999). Le risque est trop important, en l'état actuel des outils, d'obtenir un résultat à la fois

trop bruité (des extraits ont été identifiés comme discours rapporté alors qu'ils ne le sont pas) et trop silencieux (des extraits n'ont pas été identifiés).

L'implication directe sur l'application logicielle est que l'utilisateur ne pourra en aucun cas intégrer ses propres textes.

L'annotation a donc été faite manuellement, par deux annotateurs indépendants dont les marquages ont été comparés et discutés en cas de divergence de point de vue. Les balises apposées au corpus se regroupent en deux catégories majeures : délimitation et catégorisation des séquences de discours rapporté d'une part, délimitation des segments introducteurs d'autre part. Toutefois, les séquences de discours rapporté se divisent en trois familles distinctes : les passages entre guillemets (PEG) qui ne délimitent pas nécessairement un discours rapporté (cf. *supra*), les discours rapportés à proprement parler (DR) et les dialogues – généralement introduits par une suite de tirets à la ligne.

Discours rapporté	**Segments introducteurs**

`<PEG>`
`…</PEG>`
encadrent les passages entre guillemets
 	
`<DR TYPE="DD|DI|DIL|MDS|DN" >`
`…</DD>`
encadrent les phénomènes de discours rapporté
l'attribut *TYPE* identifie le type de discours rapporté, *DI* pour le discours indirect, *DIL* pour le discours indirecte libre, *MDS* pour le discours second, *DN* pour le discours nominalisé
 	
`<DIALOGUE>`
`…</DIALOGUE>`
encadrent les dialogues
 	
`<PR>`
`…</PR>`
encadrent les « passages remarquables » identifiés par les annotateurs comme tels pour leur richesse et variété en phénomènes de discours rapporté, ce qui permettra de proposer à l'utilisateur des extraits contigus et de taille conséquente pour leurs activités pédagogiques
	`<Intro`
	`mod="Vb|SV|SN|SA|SADJ"`
	`disc="DD|DI">`
	`…</Intro>`
	délimitent l'introducteur
	l'attribut *mod* identifie la catégorie de l'introducteur (nominal, verbal, adjectival…)
	
	l'attribut *disc* identifie le type de discours introduit. Cette information peut paraître redondante puisque l'introducteur est relié par des identifiants au discours introduit, mais il rend la tâche d'extraction plus simple

Tab. 1 : Balisage XML du discours rapporté

Enfin, de manière à mettre en valeur certains passages riches en séquences de discours rapportés, autant en quantité qu'en diversité, ils ont été délimités par des balises `<PR>` qui permettront une extraction simplifiée. Le tableau 1 rappelle et explicite les différentes balises utilisées.

Le tableau 1 ne présente toutefois pas un attribut complémentaire et facultatif des balises `<DR>` et `<Intro>` d'identification réciproque qui permet d'extraire un discours rapporté une fois son segment introducteur localisé, et inversement. On peut ainsi envisager l'extraction de discours rapporté en fonction de propriétés du discours introducteur.

Sont également prévues, pour des évolutions ultérieures du projet, des annotations sémantiques des segments introducteurs, précisant leur valeur sémantique et/ou énonciative, *i.e.* ce qu'ils disent de l'énoncé rapporté, du locuteur du discours citant, de celui du discours cité, etc.

5. OUTILS INFORMATIQUES ET TAL

Les outils informatiques et TAL[12] ont dans ce projet une place prépondérante. Peu au stade de la constitution du corpus, certes, mais aux différents stades d'exploitation du corpus, ils sont la clef de voûte du système pédagogique.

5.1. Constitution d'activités pédagogiques

Comme nous l'avons précisé précédemment, les outils informatiques et de traitement automatique des langues ne sont pas suffisamment compétents pour permettre une annotation automatique des phénomènes de discours rapporté.

En revanche, l'annotation morphosyntaxique robuste, bien que n'atteignant pas la perfection, donne des résultats tout à fait satisfaisants depuis un certain nombre d'années déjà, comme l'a montré par exemple la campagne GRACE d'évaluation des analyseurs syntaxiques (Adda *et al.*, 1998). On notera toutefois que certains types d'erreurs régulières devront être pris en compte dans la constitution des activités pédagogiques. Ainsi, les systèmes sont rarement capables de différencier les modes indicatifs, impératifs et subjonctifs des verbes du premier groupe.

Malgré ce risque contrôlé d'erreurs, le corpus a été annoté morphosyntaxiquement (catégorie syntaxique, sous-catégorie, lemme) afin de donner la possibilité aux utilisateurs de compléter et d'affiner leurs recherches sur corpus avec les critères ainsi calculés. Une représentation syntaxique (*tagset*) a été déterminée spécifiquement d'une part pour permettre un étiquetage indépendant de la langue et d'autre part pour permettre une traduction de ces étiquettes en langage naturel, et leur compréhension par le pédagogue.

Si les outils informatiques et TAL ont finalement un rôle mineur dans la préparation du corpus, ils trouvent toutefois une place prépondérante dans l'application pédagogique, d'une part pour la génération d'activités et d'autre part pour l'évaluation des productions de l'apprenant.

Lors de la constitution d'activités pédagogiques, les outils informatiques permettent à l'enseignant d'effectuer une extraction automatique (mais contrôlée) d'extraits du corpus correspondant aux critères qu'il juge adéquat pour telle ou telle activité. La combinaison des critères descriptifs du discours rapporté, de son segment introducteur, du contexte et du contenu de ces discours (étiquetage morphosyntaxique, lemmatisation) permet à la fois d'envisager des activités générales et des activités sur des types de discours particuliers dans des contextes bien définis, laissant ainsi à l'enseignant toute liberté dans son travail et notamment la possibilité de mettre en place une démarche fondée sur l'observation en faisant explorer le corpus par les apprenants eux-mêmes. On notera que l'interface pour l'extraction des échantillons de corpus a été conçue avec une double facette : un mode de

requête simple pour l'utilisateur peu expérimenté et un mode « expert » permettant des requêtes complexes et multicritères. D'autre part, une table associative permet de présenter à l'utilisateur des termes d'ordre pédagogique plutôt que des étiquettes XML.

Une fois l'ensemble d'échantillons défini, il est alors possible de construire des activités pédagogiques en fonction de critères et de consignes définis par l'enseignant.

5.2. Evaluation et assistance auprès de l'apprenant

Lors de l'évaluation des réponses de l'apprenant, toute une panoplie d'outils informatiques et TAL peut être mise en œuvre en fonction du type d'activité. Toutefois, ceux-ci dépendent du type de réponses attendues de l'apprenant – les formes de réponses peuvent varier de la case à cocher à la rédaction libre – et du type d'aide apportée en fonction des erreurs de l'apprenant – du simple rappel de cours accompagné d'un éventuel concordancier à l'explication contextualisée tentant d'interpréter les raisons de l'erreur de l'apprenant.

Ces différents modes d'évaluation et d'assistance sont en cours de développement en collaboration avec des enseignants.

6. INTEGRATION DANS *MIRTO*

Qu'il s'agisse de l'annotation du corpus, de la constitution d'activités pédagogiques, de la correction des erreurs d'apprenants ou de l'aide contextualisée, ces procédures font appel, chacune à leur manière, aux mêmes processus informatiques ou de traitement automatique des langues. Afin de réutiliser intelligemment chacun de ces calculs, une plateforme pédagogique est développée à l'Université Stendhal depuis quelques années.

Conçue de manière modulaire et hiérarchique, la plateforme MIRTO permet à l'enseignant, suivant son niveau d'expertise, de concevoir des scenarii pédagogiques qu'il peut mettre à disposition des apprenants tout en pouvant suivre leur évolution.

En prenant la hiérarchie *bottom-up*, du plus petit module au plus gros (de gauche à droite sur la figure 2) :

- **Fonctions :**
  Les fonctions représentent le niveau le plus bas d'objets MIRTO. Elles correspondent à un processus TAL basique comme la tokenisation (découpage du texte en formes) ou la reconnaissance de langue. Etant donnée la nature technique des fonctions et leur indépendance de toute application pédagogique, ce niveau n'est pas visible par les utilisateurs.
- **Scripts :**
  Ce niveau correspond à l'application de fonctions TAL à la didactique des langues. Un script est une suite de fonctions liées par un objectif pédagogique.
  Par exemple, la génération automatique d'exercices lacunaires est considérée comme un script car elle lie les fonctions d'identification de la langue, de tokenisation, d'analyse morphologique et de création des « trous » en fonction de paramètres entrés par l'utilisateur.
  Le fonctionnement d'un script est transparent pour l'enseignant concepteur de scénarii puisqu'il lui est présenté comme une boîte à outils.

- **Activités :**
  Une activité est la mise en œuvre d'un objectif pédagogique minimal comme travailler une notion grammaticale particulière, rédiger un paragraphe sur un sujet, réviser des conjugaisons, etc. Elle se réalise donc par la construction d'un espace de travail pour l'apprenant, lui permettant d'atteindre le but visé. Elle correspond à ce que l'on désigne traditionnellement comme exercice. À un niveau purement didactique, la conception d'activités est une tâche opérée par les enseignants de langue à l'aide d'un outil spécifique de MIRTO : l'éditeur d'activités. L'éditeur d'activités est un environnement de conception (système auteur) permettant de visualiser et de manipuler des objets et des outils pédagogiques tels que des textes (ou corpus de textes), des scripts et des consignes.
- **Scenarii :**
  Ce niveau permet aux enseignants de définir des séquences d'activités. Les scénarii permettent de répondre à leurs objectifs pédagogiques en fonction du parcours de l'apprenant.
  Ce dernier peut être différent pour chaque apprenant. En effet, chacun d'entre eux aura un processus d'apprentissage personnel lié à différents facteurs. MIRTO prend en compte cette réalité en proposant la création de scénarii non linéaires. La parcours du scénario dépendra du processus individuel de chaque apprenant (parcours d'apprentissage, évaluation...). Celui-ci sera stocké dans un historique propre à chaque apprenant.
  Selon son parcours dans un scénario donné, un apprenant pourra être redirigé vers une activité de remédiation, retenter la même activité sur autre texte ou encore simplement avancer dans le scénario.

Dans le cas de l'utilisation de la plateforme MIRTO pour l'enseignement et l'apprentissage du discours rapporté on aura recours notamment :

- aux fonctions
  - d'étiquetage morphosyntaxique,
  - d'extraction et de concordance de contextes selon des critères combinant les propriétés des discours rapportés, les traits syntaxiques ou des lemmes des verbes des introducteurs, etc.
- aux scripts
  - de création d'exercices lacunaire paramétrés par exemple pour permettre à l'apprenant de déterminer le temps et le mode des verbes introducteurs ;
  - de création d'exercices de cases à cocher afin que l'apprenant détermine par exemple le type du discours rapporté ;
  - de génération d'exemplier pour aider l'apprenant en fonction des erreurs constatées.

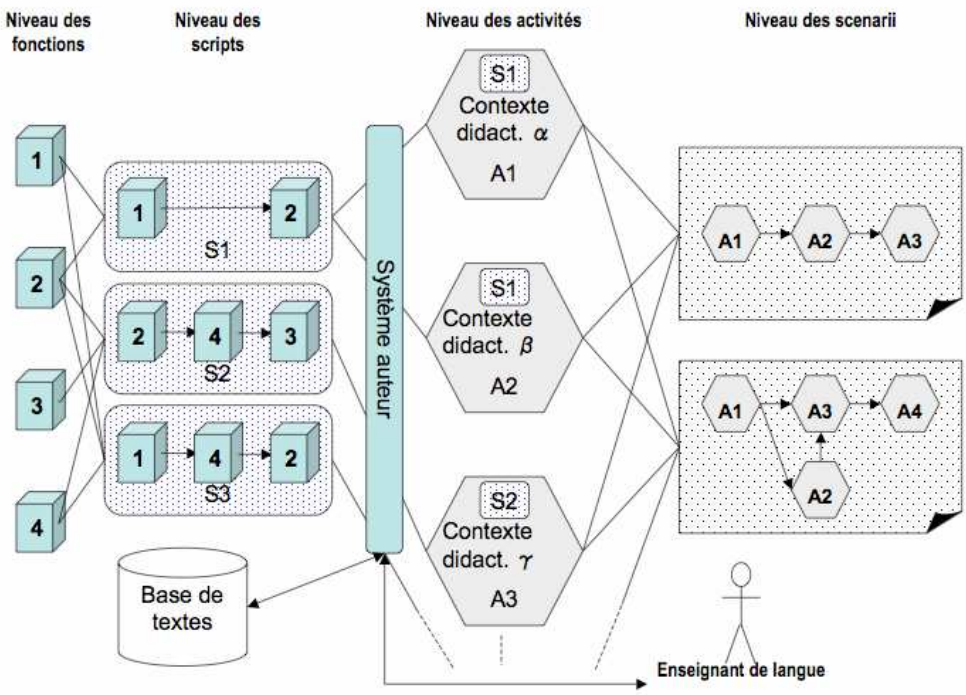


Fig. 2 : Principe modulaire de la plate-forme MIRTO

7. EXPLOITATION PEDAGOGIQUE

Les propositions d'exploitation pédagogique du corpus que nous présenterons ici ont été conçues en collaboration avec des enseignants et des formateurs ; il s'agit d'une part d'exercices relatifs au discours rapporté intégrés dans Mirto et s'adressant à deux publics différents : un public de FLE niveau intermédiaire à avancé et un public de collégiens. D'autre part, on rendra compte d'une exploitation pédagogique « off line » menée avec un public d'étudiants du supérieur sur la question des guillemets.

7.1. Exercices intégrés dans Mirto

À côté des scripts développés pour permettre à l'enseignant d'interroger facilement le corpus, et d'en extraire des exemples ou des passages intéressants, ainsi que des informations sur la fréquence des éléments balisés, les activités générées automatiquement à partir du corpus consistent en des exercices d'identification, des exercices lacunaires et des exercices de type case à cocher. Aux exercices traditionnels de manipulations morphosyntaxiques du type transformation du discours direct en discours indirect, ont été mises en place des activités concernant le repérage du discours citant, du discours cité et des segments introducteurs, leur manipulation et l'analyse grammaticale, sémantique ou énonciative de ces segments, des différents types du discours rapporté et des passages entre guillemets.

Pour en donner quelques exemples, nous pouvons citer d'abord un exercice d'identification portant sur les segments introducteurs de discours rapporté et de leur

nature syntaxique : la consigne est de surligner ce qui dans le discours cité annonce le discours citant, en recourant à un code couleur rendant compte de la catégorie syntaxique de cet élément introducteur.

- C'était si bien imité que je ne pus m'empêcher de rire. Mais lui, **très froidement** :
- Vous croyez que je plaisante... regardez mes yeux[13].

Ce travail sur les segments introducteurs mobilise des compétences grammaticales, sur la nature syntaxique des segments, et il favorise le repérage de la diversité des modes d'introduction du discours rapporté et des effets sémantiques induits par ces modes d'introduction.

On retrouve cet objectif de sensibilisation à la variété des modes d'introduction et aux effets qu'ils produisent dans divers exercices, comme celui consistant à combler des lacunes dans les segments introducteurs de discours rapporté : l'apprenant doit choisir et conjuguer le verbe qui lui semble le plus approprié parmi une liste proposée. Cette consigne implique qu'il fasse des inférences à partir du contexte pour le choix sémantique du verbe, outre les compétences grammaticales en production et réception requises par la conjugaison et la gestion de la concordance de temps. On peut citer aussi le cas d'un exercice lexico-sémantique de type case à cocher, qui touche à la dimension pragmatique des effets de sens induits par les segments introducteurs : les apprenants doivent déterminer, à partir d'énoncés dont les segments introducteurs sont soulignés, quel acte de langage est spécifié par ces segments (*e.g.* requête, refus, jugement, etc.). D'autres exercices du même type peuvent porter sur l'attitude du locuteur exprimée par le segment introducteur ou encore sur la fonction même du segment introducteur : sert-il à spécifier l'acte de langage, exprime-t-il l'attitude du locuteur, concerne-t-il l'accomplissement de l'acte phatique ?

Une autre série d'exercices s'adresse à un public plus avancé en FLE ou à un public de lycéens en FLM, voire d'étudiants, et vise le développement de compétences métalinguistiques par l'identification de la fonction des guillemets, des types de discours rapporté ou encore des types de discours autres, comme un exercice de type case à cocher où l'apprenant doit analyser dans un énoncé à quel type de phénomène il a affaire : par exemple, discours indirect libre, discours narrativisé, ou modalisation en discours second, dans

> Bientôt tout l'îlot apparut, comme s'il fût descendu d'un nuage ; puis, la mer se montra suivant un plan circulaire, infinie dans l'est, mais bornée dans l'ouest par une côte élevée et abrupte. **Oui ! la terre était là**. Là, le salut, provisoirement assuré, du moins[14].

La complexité de ces phénomènes de discours autres ou de mise entre guillemets est telle qu'il est des cas où plusieurs pistes interprétatives sont envisageables. L'intérêt de ces exercices tient alors certes à l'identification de ces interprétations, et il est donc prévu des doubles réponses, mais aussi à la discussion qu'ils peuvent susciter en classe sur cette pluralité interprétative, et sur les limites des typologies linguistiques de ces phénomènes. L'importance de ces prolongements didactiques est vraie également au niveau de la correction des exercices : si celle-ci peut parfois être entièrement automatisée, dans les exercices sur la valeur sémantique des verbes ou des segments introducteurs, là encore le choix à faire n'est pas nécessairement univoque, et la correction appelle à être complétée par une discussion en classe sur les divers effets induits par le choix des segments introducteurs.

7.2. Exploitation « off line » du corpus

Le seul accès au corpus enrichi fournit d'autres possibilités d'exploitations didactiques. Nous avons ainsi expérimenté une exploitation « off line » du corpus, auprès d'étudiants de première année de sciences du langage, dans le cadre d'un cours de linguistique de l'énonciation. L'objectif est de sensibiliser les étudiants au phénomène des guillemets, pour leur faire élaborer une typologie de leurs fonctions.

Les étudiants travaillent par binômes et explorent un des textes du corpus, un roman, à partir d'une commande de détection automatique des passages entre guillemets. Ils ont pour consigne d'analyser la fonction des guillemets dans chaque passage, tout en repérant dans leur entourage d'éventuels indices de cette fonction. À partir de ces analyses, chaque binôme établit une typologie des guillemets, confrontée alors à celles proposées dans les autres groupes.

Il s'agit donc de privilégier une démarche inductive, fondée sur l'observation en corpus, et relayée par des discussions en grands groupes avec pour objectif le développement d'une réflexion métalinguistique. Le rôle de l'enseignant est très important au cours de la séance, ses interactions permettant aux étudiants d'approfondir leurs analyses en corpus au sein des binômes et lors de la discussion finale des différentes fonctions des guillemets. Les ressources développées sont autant d'outils favorables à ce type de démarche d'analyse, en offrant un accès rapide à un nombre élevé de passages entre guillemets et la possibilité pour l'enseignant de se référer aux annotations quant à la fonction des guillemets dans ces passages et à leur éventuelle polysémie.

8. CONCLUSION

Nous avons présenté ici les premiers résultats d'un travail collaboratif inter-disciplinaire alliant linguistes, talistes et didacticiens. La constitution d'une ressource linguistique pour l'enseignement/apprentissage du discours rapporté a permis de confirmer et d'affiner la typologie prise en considération initialement. De plus, l'annotation manuelle des phénomènes énonciatifs complétée d'une annotation morphosyntaxique automatique a rendu possible l'intégration de ce corpus au sein de la plateforme d'apprentissage des langues préexistante, MIRTO. Les premières activités pédagogiques portant sur les phénomènes de manifestation de la subjectivité du locuteur ont été développées sur la plateforme et utilisées en cours auprès d'apprenants du français langue étrangère (FLE), de collégiens et d'étudiants en sciences du langage.

S'il convient de finaliser la conception des scenarii pédagogiques, ces premières expériences attestent de l'intérêt d'une mise à disposition plus large du corpus, qui permette aux enseignants d'exprimer leur créativité pédagogique grâce à la modularité de MIRTO.

9. RÉFÉRENCES

G. Adda, J. Lecomte, J. Mariani, P. Paroubek, M. Rajman, 'The GRACE French Part-of-Speech Tagging Evaluation Task', in *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Granada, mai 1998

Antoniadis G., Echinard S., Kraif O., Lebarbé T., Loiseau M., Ponton C. 2004. 'NLP-based scripting for CALL activities', in *COLING 04*, août 2004. Genève, Suisse.

Authier-Revuz J. 1979. 'Problèmes posés par le traitement du discours rapporté dans une grammaire de phrases', *Linguisticae investigationes*, 3-2, p. 211-228.

Authier-Revuz J. 1992-93. 'Repères dans le champ du discours rapporté' (I) et (II). *L'information grammaticale,* 55 et 56, p. 38-42 et p. 10-15.
Boch F. & Grossmann F. 2001. (ed.) *Apprendre à citer le discours d'autrui*. *Lidil* 24. Grenoble : Université Stendhal.
Bourion E. & Malrieu D. 2004. 'Types de discours rapportés et typologie des narrateurs dans le genre romanesque', Colloque *Dans la jungle des discours*, Cadiz, 11-13 mars 2004.
Bres J. & Verine B. 2002. 'Le bruissement des voix dans le discours : dialogisme et discours rapporté', *Faits de Langues,* 19, p. 159-169.
Maingueneau D. 1994. *L'énonciation en linguistique française*. Paris : Hachette.
Kerbrat-Orecchioni C. 1997. *L'énonciation : de la subjectivité dans le langage*. Paris : Seuil (3ème éd., édition originale 1980)
Mourad G. & Desclés J.P. 2002. 'Citation textuelle : identification automatique par Exploration Contextuelle', *Faits de Langues*, 19 : Le Discours rapporté, p. 179-89.
Rabatel A. 2003. 'Les verbes de perception en contexte d'effacement énonciatif : du point de vue représenté aux discours représentés', *Travaux de linguistique*, 46-1, p. 49-88.
Rosier L. 1999. *Le discours rapporté. Histoire, théories, pratiques*. Paris, Bruxelles : De Boeck-Duculot.

10. NOTES

1. Par « talistes », nous entendons informaticiens-linguistes, chercheurs en traitement automatique des langues (TAL)
2. Notons toutefois que G. Mourad et J.P. Desclés (2002) travaillent à la reconnaissance automatique du discours direct, et que E. Bourion et D. Malrieu (2004) cherchent à analyser les types de discours rapporté comme indices de modes narratoriaux.
3. J. Verne, *L'île mystérieuse*.
4. A. Daudet, *Lettres de mon moulin*.
5. *Ibidem*.
6. Cf. cet extrait de A. Daudet, *Lettres de mon moulin* :
Le petit bois de chênes verts semble lui faire signe :
- Venez donc par ici, monsieur le sous-préfet; pour composer votre discours, vous serez bien mieux sous mes arbres...
7. A. Daudet, *Lettres de mon moulin*.
8. G. Dupuy, « Les grandes manœuvres américaines », *L'Express*, 12/02/2002.
9. A. Daudet, *Lettres de mon moulin*.
10. Par œuvre récente, nous entendons les œuvres pour lesquelles les auteurs ou les ayant droit ont encore un droit patrimonial (notamment la liberté de choisir la diffusion à titre gratuit ou payant de leurs œuvres). Afin de déterminer si une œuvre est « tombée dans le domaine public », les calculs ne s'arrêtent pas au simple principe des 70 ans après l'année civile qui suit le décès de l'auteur, mais intègrent aussi notamment les périodes de guerre pendant lesquelles les auteurs n'ont pas pu faire valoir leurs droits.
11. Contrairement à certaines croyances, les créations littéraires et artistiques mises à disposition sur Internet ne sont pas systématiquement libres de droit. Certains auteurs ont refusé, pour des raisons qui leur sont propres, l'utilisation et la mise à disposition de leurs écrits dans la plateforme pédagogique présentée ici, ce que nous avons respecté.
12. Une interprétation de la notion d'outil de traitement automatique des langues pourrait être « outil manipulant des données langagières ». Nous y préférerons la définition suivante : outils manipulant des données langagières intégrant implicitement ou explicitement des connaissances linguistiques comme ressources pour les traitements.
13. A. Daudet, *Lettres de mon moulin*.
14. J. Verne, L'île mystérieuse.

THÉMATIQUE LINGUISTIQUE

QUELQUES ADVERBES MODALISATEURS DANS LE FRANÇAIS PARLÉ L2 ET L1 : INSERTION DANS LE *PARAGRAPHE ORAL*

Victorine Hancock
Université de Stockholm

1. INTRODUCTION

1.1. But et questions de recherches

L'étude vise à examiner l'usage de quelques adverbes récurrents dans le français parlé des locuteurs natifs (LN) et des locuteurs suédophones en français L2 (*i.e.* locuteurs non natifs, LNN). Le but général de cette étude essentiellement descriptive est de trouver des critères qui puissent caractériser l'organisation discursive des apprenants avancés (voir Bartning, 1997 et 2004 pour une description de l'apprenant avancé et du projet *InterFra*[1]). Dans l'étude nous considérerons la position des adverbes dans le *paragraphe oral* (Morel & Danon-Boileau, 1998), ce dernier étant déterminé par la structure informationnelle et par des traits intonatifs suprasegmentaux (voir section 3 ci-dessous). Conway (2005), dans sa thèse sur le paragraphe oral, a remarqué que chez les deux apprenantes (avancées) de français L2 étudiées, les éléments modaux du *préambule* du paragraphe oral (partie thématique et modale) sont moins fréquents que chez les locuteurs natifs. Les questions principales que nous nous posons ici sont les suivantes : quelles sont les adverbes de phrase modalisateurs les plus fréquents chez les LNN et les LN ? Y a-t-il une différence entre les deux groupes de locuteurs concernant les positions occupées par les adverbes dans le paragraphe oral ?

1.2. Un modèle pour segmenter la chaîne parlée

Pour analyser la chaîne parlée en segments pertinents nous nous appuyons sur le modèle du *paragraphe oral* proposé par Morel & Danon-Boileau (1998), ainsi que sur les hypothèses macro-syntaxiques générales proposées par Blanche-Benveniste (1997) et Berrendonner (1990). Selon Morel & Danon-Boileau (1998), qui partent de la séparation classique du *modus* et de *dictum* (Bally 1944), le français parlé obéit au principe de *décondensation*, à savoir qu'un certain nombre de constituants discursifs sont dissociés de la construction verbale réalisée par le *rhème*. Le paragraphe oral consiste, au moins, en un *préambule* et un *rhème.* Chez le locuteur natif de français, le *préambule* peut souvent être très décondensé. Dans le paragraphe oral « *de toute façon moi je crois ce qui est étranger / ça fait toujours peur quoi* » (exemple du Corpus *InterFra*, locuteur natif), le préambule est constitué de quatre éléments : *ligateur, point de vue, modus dissocié, support lexical disjoint.*

Dans la section (3) ci-dessous nous allons donner avec davantage de détails, les critères du paragraphe oral, afin de pouvoir l'identifier avec ses constituants dans la chaîne parlée.

2. CORPUS ET MÉTHODE

Les locuteurs de cette étude sont dix étudiants suédophones (24 523 mots), ainsi que dix locuteurs natifs de français (27 686 mots), qui contribuent chacun avec une interview (voir Tableau 1). Nos données sont tirées du corpus *InterFra* (Bartning, 1997) constitué à l'université de Stockholm en Suède. Le corpus comprend des enregistrements d'étudiants suédophones, ainsi que de locuteurs natifs francophones, dont les derniers sont des étudiants en échange en Suède. Le thème des interviews est centré autour des études, des loisirs, des projets d'avenir. Les étudiants suédophones proviennent de deux groupes d'apprenants avancés du corpus : six futurs enseignants de français (âge 23-34 ans) et quatre étudiants du cursus général (âge 19-26 ans). Selon les stades d'acquisition proposés par Bartning & Schlyter (2004), les informants appartiennent aux niveaux 4 (Yvonne), 5 (Anders, Anita, Lena, Malena, Marie, Mona) et 6 (Ida, Kerstin, Pernilla). Ces niveaux correspondent aux stades avancés (bas, moyen et supérieur).

Une analyse acoustique des contours intonatifs de notre échantillon a été faite à l'aide du logiciel WinPitch. À partir des contours, les variations de F0 (la fréquence du fondamental de la voix) et d'I (l'intensité) nous ont servie d'indices pour segmenter la chaîne parlée en paragraphes oraux.

En ce qui concerne l'analyse quantitative, le test de chi-carré a été appliqué pour savoir s'il y a des différences significatives entre LNN et LN.

LNN	Ani	Ida	Ker	Mal	Mon	And	Yvo	Mar	Per	Len	**Total**
	2150	2612	2118	2763	2644	2929	2307	1837	2063	3100	**24 523**
LN	Ann	Eri	Gaë	Hen	Isa	Kar	Lau	Mél	Oli	Pie	**Total**
	3622	1792	3342	2068	2413	1938	3035	4129	1991	3356	**27 686**

Tableau 1: N de mots dans les interviews

3. DÉFINITIONS : LES INDICES D'IDENTIFICATION DU PARAGRAPHE ORAL

Résumons ici les indices les plus centraux pour identifier le paragraphe et dont nous nous sommes servies pour notre analyse[2]. Les indices se situent à deux niveaux, dont l'un concerne les constituants et les sous-constituants du paragraphe (*indices lexicaux et syntaxiques*). L'autre niveau concerne *l'ensemble* des segments – préambules et rhèmes – qui constituent le paragraphe (traits *suprasegmentaux*). Ce dernier niveau repose sur des *indices intonatifs*.

3.1. Indices lexicaux et syntaxiques

Pour identifier les constituants (préambule(s), sous-constituants du préambule et rhème(s)) d'un paragraphe, nous pouvons recourir à des indices lexicaux et syntaxiques. Voici un paragraphe contenant un préambule composé d'éléments discursifs typiques[3] :

	préambule	*rhème*	*postrhème*
(1)	oui parce que : il me semble ici ce que ça voudrait représenter ↑	c'est les amandes	ça ↓

Le préambule ci-dessus se décompose en sous-éléments discursifs de la façon suivante :

- ligateur : *oui*
- ligateur : *parce que*
- point de vue : *moi*
- modus dissocié : *il me semble*
- cadre spatial : *ici*
- support lexical disjoint (coréférentiel avec le sujet du rhème) : *ce que ça voudrait représenter*
  La structure générale du *rhème* peut se résumer ainsi :
- [c'est / il y a / pronom + V conj] + *x* (+ ponctuant) où *x* est un SN, SP ou adverbe : *c'est les amandes*

Le *postrhème* constitue le troisième segment du paragraphe : *ça*. Le postrhème est soit un élément nominal coréférent avec le sujet du rhème, soit un élément modal.

3.2. Indices intonatifs de démarcation du paragraphe (traits suprasegmentaux)

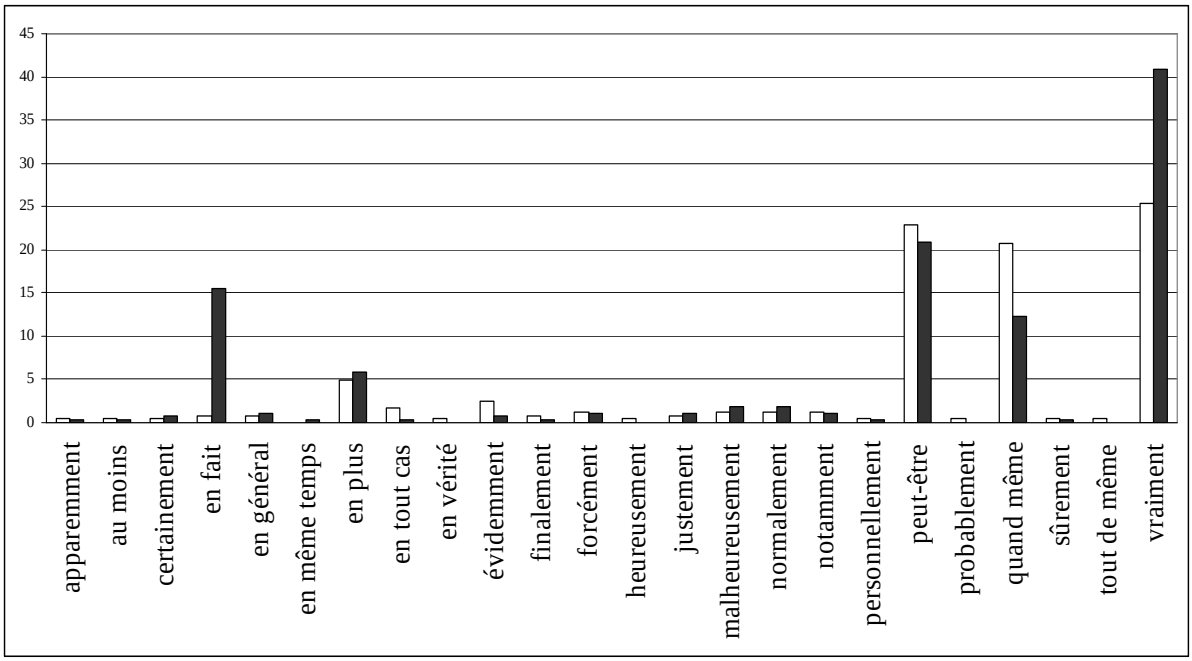
Pour l'analyse de la section (3), les indices permettant de découper la chaîne parlée en paragraphes oraux sont les suivants : 1) à la fin du paragraphe, il y a une chute conjointe de F0 (la fréquence du fondamental de la voix) et de I (l'intensité), sans allongement de la syllabe finale. 2) Une remontée de F0, sur la syllabe finale d'un constituant ou plusieurs constituants, sert à *unifier / recatégoriser* rétroactivement cet ensemble en préambule pour le discours qui suit, et signifie que le paragraphe continue. 3) La présence d'un postrhème (voir ci-dessus) indique la fin du paragraphe. Le postrhème suit immédiatement le rhème (sans pause entre les deux segments) et est doté d'une F0 plus basse et une I plus faible que le rhème qui précède.

4. ANALYSE

4.1. Un premier regard sur les adverbes modalisateurs du corpus

Dans une étude récente (Hancock, 2005) sur le *modus dissocié* du paragraphe oral (voir section 4.1 ci-dessus) chez six apprenants avancés du corpus *InterFra*, nous avons observé qu'un certain nombre d'adverbes de phrase en *–ment*, comme *justement*, *évidemment* et *forcément*, sont très rares ou absents chez les LNN. Ces derniers adverbes se trouvent parmi les cinq adverbes en *–ment* les plus fréquents dans le français parlé (*le corpus de référence de français parlé*, Bilger, 2004).

Un premier regard (Graphique 1) sur la fréquence des adverbes (susceptibles d'être des) modalisateurs de phrase chez les LNN et les LN de la présente étude, montre que trois adverbes sont beaucoup plus fréquents (10-50 fois) que la majorité des autres, à savoir *vraiment*, *peut-être* et *quand même*. Ces trois adverbes sont les seuls à être employés par l'ensemble des informants du corpus (LNN et LN). Les seuls adverbes concurrents avec les trois adverbes en tête sont *en fait* (employé seulement par 7 LN) et *en plus*.



Graphique 1 : Adverbes modalisateurs / 10 000 mots. Blanc : LNN; Noir : LN.

Le nombre d'occurrences ramenées à 1/10 000 pour *vraiment*, *peut-être* et *quand même* est 25, 23 et 21 respectivement, pour les LNN (Tableau 2). Nous pouvons constater que le nombre d'occurrences des adverbes chez les LN ressemble de façon étonnante à celui des LNN (41, 21 et 12 chez les LN). Dans le Tableau 2, les chiffres observés des deux groupes de locuteurs sont rassemblés avec ceux de Bilger (2004) et du corpus Orléans et de Bristol (Beeching, 2005). Il est difficile de tirer des conclusions fines à partir de la comparaison entre les corpus du fait que la fréquence absolue des adverbes est basse. Toutefois, si on compare les trois corpus pour *vraiment* et *quand-même*, constatons que les chiffres sont du même ordre. Le sous-corpus de référence qui ressemble, par des critères extérieurs, le plus au nôtre est le corpus « parole privée », qui « correspond à des entretiens dans lesquels les locuteurs racontent un événement de leur vie ou présentent un savoir professionnel ou autre » (Bilger, 2004, p. 63). Nous pouvons observer que ce dernier sous-corpus, parmi les trois sous-corpus, ressemble aussi le plus au corpus *InterFra*, quant au nombre d'occurrences pour *vraiment* (16/10 000 mots). *Vraiment* est dans le sous-corpus « privée » de loin l'adverbe en *–ment* le plus fréquent (trois fois plus fréquent que celui en deuxième position, Bilger, 2004, p. 66).

		VRAIMENT	PEUT-ÊTRE	QUAND-MÊME
Corpus *InterFra* :	LNN	25	23	21
	LN	41	21	12
Corpus d'Aix (Bilger 2004) trois sous-corpus	privée	16		
	professionnelle	11		
	publique	8		
Beeching (2005)	Corpus Orléans			14
	Corpus Bristol			23

Tableau 2 : N d'occurrences / 10 000 mots

Après ce premier regard quantitatif sur les adverbes modalisateurs, nous allons regarder de plus près *vraiment*, *peut-être* et *quand-même*, les trois adverbes les plus fréquents chez les LNN et employés par tous les informants du corpus.

4.2. Vraiment, peut-être et quand-même

Dans cette section, nous allons voir quelles positions les trois adverbes occupent dans le paragraphe oral et quelle est la répartition des positions dans les deux groupes de locuteurs. Voici ci-dessous des exemples typiques de chacune des positions que peuvent occuper les adverbes. Les adverbes examinés peuvent occuper, à part une place dans le *rhème* (**rh**), une place soit dans le préambule (**pr**) – comme *modus dissocié* – soit à la fin du paragraphe comme *postrhème* (**po**) :

4.2.1. VRAIMENT

(2a) **pr :** je crois qu' **rh :** il y a ***vraiment*** les les deux versants les deux côtés (Ann, LN)

(2b) **pr :** ***vraiment*** **rh :** ils se remettent en cause (Mél, LN)

(2c) **pr :** oui **rh :** on a mis beaucoup de temps sur euh ces devoirs **po :** ***vraiment*** (And, LNN)

La position la plus fréquente des trois positions est celle dans le rhème (Ex. 2a). Cependant, une importante proportion des LN (7 informants), l'emploient aussi dans le préambule (Ex. 2b). *Vraiment* est peu employé comme postrhème (Ex. 2c) chez les LN (2 informants), mais davantage par les LNN (8 informants).

4.2.2. PEUT-ETRE

(3a) **pr :** et donc euh **rh :** c'est plus intéressant pour eux ***peut-être*** de voir comment on apprend le français (Ida, LNN)

(3b) **pr :** enfin bon ***peut-être*** aussi qu' **rh :** on est entre étudiants et tout ça (Gaë, LN)

(3c) **pr :** le plus difficile pour moi c'est / en cette période de l'année **rh :** c'est le manque de soleil **po :** ***peut-être*** (Hen, LN)

Dans les deux groupes de locuteurs, la position dans le rhème (Ex. 3a) est la plus fréquente, suivie de la position dans le préambule (Ex. 3b) et en troisième, comme postrhème (Ex. 3c). Il est intéressant de noter que, chez les LNN, bien que *peut-être* soit placé devant le rhème (préambule), il est souvent « intégré » intonativement dans le rhème, comme le montre les exemples suivants (Ex. 4a-c) chez trois LNN :

(4a) I : mhm / les contacts ?
E : les contacts / ***peut-être*** il faut faire un effort (And, LNN)

(4b) ***peut-être*** on aura aussi beaucoup plus de : (Len, LNN)

(4c) ou ***peut-être*** c’est pas toutes (Yvo, LNN)

4.2.3. QUAND-MEME

(5a) **pr :** mais **rh :** c'est ***quand même*** un peu gras la nourriture (Mal, LNN)

(5b) **pr :** parce que je crois ***quand même*** qu'on parle / dans la vie quotidienne **rh :** on parle des même choses tout le temps (Per, LNN)

(5c) **pr :** les frites **rh :** c'est c'est bon **po :** ***quand même*** (Ker, LNN)

(5d) **pr :** oui **rh :** ça leur coûte un peu plus cher que quand j'étais à Lille **po :** ***quand même*** (Kar, LN).

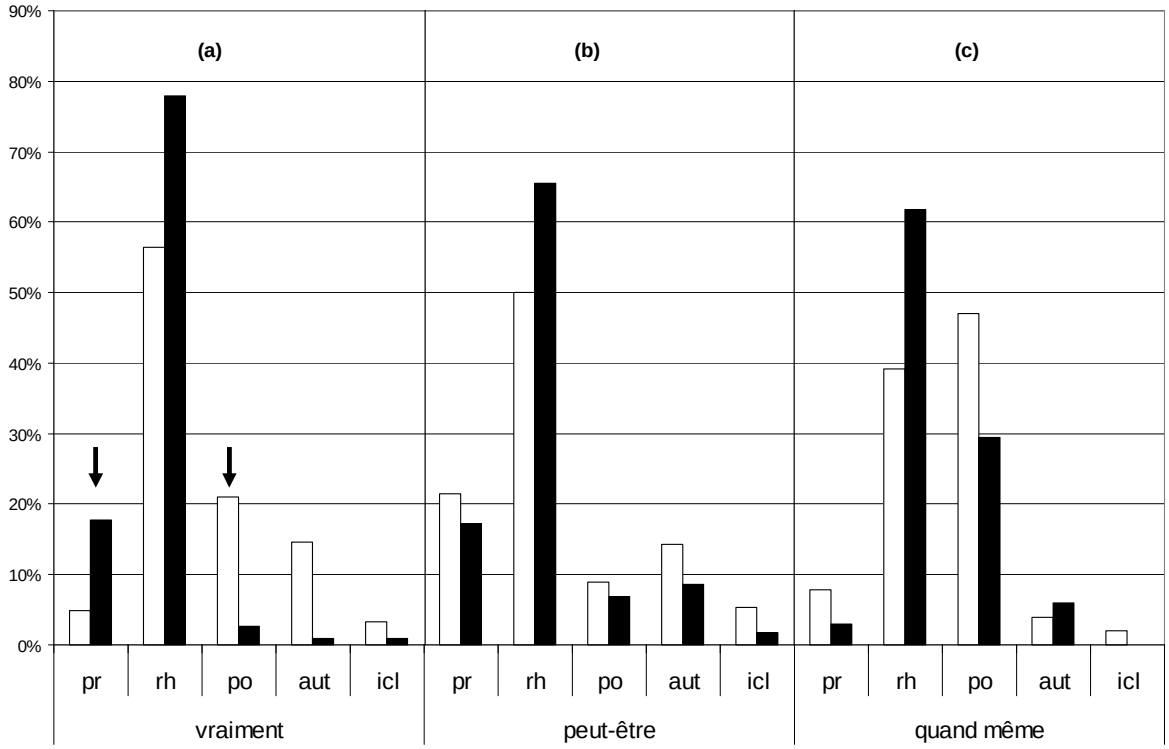
La position la plus fréquente de *quand-même* chez les LN est dans le rhème (Ex. 5a), mais pour les LNN, la position comme postrhème (Ex. 5c-d) prime. Dans les deux groupes de locuteurs, la position dans le préambule (5b) est la moins fréquente.

4.2.4. CONCLUSION DE L'ANALYSE QUANTITATIVE EN POSITIONS Les adverbes vraiment, peut-être et quand-même sont majoritairement employés dans le rhème par les locuteurs natifs (LN), ce qui est le cas aussi pour vraiment et peut-être chez les locuteurs non natifs (LNN) (Graphique 2a-c, ci-dessous). En ce qui concerne quand même, c'est la position à la fin du paragraphe qui domine chez les LNN (Graphique 2c).

La position dans le préambule de *vraiment* est plus fréquente chez les LN que chez les LNN (LN : 18 % vs. LNN : 5 % ; différence significative : $\chi 2 = 6.4$ et $p < 0.05$*) (Graphique 2a). Ces résultats pourraient être comparés avec ceux de Bilger (2004), qui montre que dans le corpus « privée » on relève seulement 6 % des occurrences de *vraiment* à l'initiale. Cependant, l'auteur réfère à un corpus constitué d'entrevues d'hommes politiques, dans lequel la place initiale de *vraiment* correspond à 15 % des emplois (Bilger, 2004, p. 76 note 8). Il nous semble que cette position reflète un registre « affectif », apparemment peu employé par les LNN.

Pour les trois adverbes, la position comme postrhème est plus fréquente chez les LNN que chez les LN (la différence est significative pour *vraiment* : $\chi 2 = 16.6$ et $p < 0.001$*** et pour *quand même* : $\chi 2 = 5.4$ et $p < 0.05$*)[4] (Graphique 2a-c).

Les autres tendances relevées sont les suivantes : chez les LNN on trouve une fréquence élevée de *vraiment* et *peut-être* en tant que rhèmes « autonomes » comme *pas vraiment, oui peut-être* (Graphique 2a-b). En ce qui concerne *peut-être,* les différences entre LN et LNN se trouvent plutôt au niveau de l'intonation : les LNN ont tendance à placer *peut-être* en tête du rhème, intégré intonativement dans le rhème.



Graphique 2 (Blanc : LNN ; noir : LN,

pr = préambule, rh = rhème, po = postrhème, aut = autonome, icl = inclassable)

5. CONCLUSION

En guise de conclusion, bien que le comportement des trois adverbes *vraiment*, *peut-être* et *quand même* diffère, il semblerait que les LNN avancés ont tendance à déplacer les adverbes examinés du début du paragraphe, dans le rhème (*vraiment*), mais aussi du rhème dans le postrhème (*vraiment, quand même*). Nous avons aussi observé une tendance chez les LN à « autonomiser » les adverbes *vraiment* et *peut-être*. Pour trouver des hypothèses explicatives du comportement des adverbes il faudrait regarder de plus près la relation entre forme et fonctionnement.

6. RÉFÉRENCES

Bally, **C.** 1944. *Linguistique générale et linguistique française*. Berne : A. Francke.

Bartning, **I.** 1997. « L'apprenant dit avancé et son acquisition d'une langue étrangère – tour d'horizon et esquisse de la variété avancée ». *A.I.L.E. (Acquisition et interaction en langue étrangère) 9.* Paris : DEPA – Université Paris VIII, p. 9-50.

Bartning, I. & Schlyter, S. 2004. « Itinéraires acquisitionnels et stades de développement en français L2 ». *French Language Studies*, 14, p. 281-299.

Beeching, K. 2005. « Politeness-induced semantic change : The case of "quand même" ». *Language Variation and Change 17*, p. 155-180.

Berrendonner, A. 1990. « Pour une macro-syntaxe ». *Travaux de linguistique*, 21, p. 25-31.

Bilger, M. 2004. « Quelques données sur les adverbes en ment dans le corpus de référence de français parlé ». *Recherches sur le français parlé*, 18, p. 63-82.

Blanche-Benveniste, C. 1997. *Approches de la langue parlée en français*. Paris : Ophrys.

Conway, Å. 2005. *Le paragraphe oral en français L1, en suédois L1 et en français L2*. Université de Lund. Études Romanes de Lund 73. Thèse de doctorat.

Hancock, V. 2000. *Quelques connecteurs et modalisateurs dans le français parlé des apprenants avancés. Étude comparative entre suédophones et locuteurs natifs*. Thèse de doctorat. Département de français et d'italien, Université de Stockholm, Suède.

Hancock, V. 2005. « Étude intonodiscursive des séquences introduites par "parce que" dans le français parlé des apprenants avancés suédophones ». *Actes du colloque international en sciences du language : APLIC (acquisition, pratiques langagières, interactions et contacts), juin 25-26 2004*, Paris III-Nouvelle Sorbonne.

Hancock, V. & Kirchmeyer, N. 2004. « Discourse structuring in advanced L2 French : The relative clause », dans J.-M. Dewaele (ed.). *Focus on French as a Foreign Language*. Multilingual Matters : Clevedon, England.

Kirchmeyer, N. 2002. *Étude de la compétence textuelle des lectes d'apprenants avancés. Aspects structurels, fonctionnels et informationnels*. Thèse de doctorat. Département de français et d'italien, Université de Stockholm, Suède.

Lacheret-Dujour, A. & Beaugendre, F. 1999. *La prosodie du français*. Paris : CNRS éditions.

Morel, M.-A. & Danon-Boileau, L. 1998. *Grammaire de l'intonation. L'exemple du français. Bibliothèque de Faits de Langues*. Paris : Ophrys

7. NOTES

1. Voir aussi Hancock, 2000 ; Kirchmeyer, 2002 ; Hancock & Kirchmeyer, 2004, pour d'autres tentatives de description de l'organisation discursive de l'apprenant avancé, dans le cadre du projet *InterFra*.
2. Pour des précisions concernant l'analyse en paragraphes oraux voir aussi, par exemple, Lacheret-Dujour & Beaugendre, 1999 et Conway, 2005.
3. L'exemple de Lorenza Mondada (Stockholm dec. 2005) : le locuteur est un agronome dans un atelier de travail devant une représentation cartographique d'une ferme.
4. Étant donné que le nombre d'occurrences de *quand même* de 10 informants était trop bas pour vérifier l'hypothèse nulle, nous avons ajouté 5 informants avancés du corpus *InterFra* à chaque groupe de locuteurs (15 locuteurs au total / groupe), ce qui a donné une différence significative.

ANNOTATION DE LA CORÉFÉRENCE ENTRE EXPRESSIONS RÉFÉRENTIELLES

Olivier Tardif
France Télécom, DELIC (U. de Provence)

1. INTRODUCTION

On dit de deux expressions qu'elles sont coréférentes lorsqu'elles désignent la même entité :

> **Jean-Pierre Raffarin** s'est présenté hier matin devant les députés UMP. Pour sa première rencontre de la rentrée avec le groupe majoritaire, **le premier ministre** n'était pas venu débattre de ses projets. *(*Le Figaro *07/10/04)*

Les expressions coréférentes peuvent également être constituées de pronoms : dans l'exemple ci-dessus, on pourrait en effet remplacer ***le premier ministre*** par ***il***. On considère ces pronoms comme des anaphores, *i.e.* comme des expressions nominales référant à d'autres expressions et dont l'interprétation dépend de celles-ci (Huang, 2000).

Nombreux sont les travaux portant sur la résolution automatique des liens de coréférence : d'une part le problème est complexe, faisant intervenir plusieurs niveaux d'interprétation (morphologie, syntaxe, sémantique, pragmatique), et d'autre part la plupart des applications en TALN profiteraient de telles informations. On peut diviser en deux classes les algorithmes décrits dans le cadres de ces travaux : les méthodes symboliques, incluant les systèmes à base de règles, de calcul de similarité ou les systèmes de contraintes/préférence ; et les méthodes statistiques basées sur des calculs de probabilité ou par arbres de décision.

Le présent travail se divise comme suit : premièrement nous décrirons un corpus de textes journalistiques mis sur pied en vue d'une annotation des relations de coréférence ; ensuite nous décrirons le schéma d'annotation utilisé ; et finalement, nous décrirons quelques données statistiques que cette annotation nous a permis d'obtenir, qui montrent certaines limites d'une approche purement symbolique du problème.

2. CONSTITUTION DU CORPUS

Lors d'une étude préliminaire portant sur quelques textes choisis au hasard dans le corpus du journal Le Monde, nous avons constaté une variation entre les différents types d'articles (courrier des lecteurs, brèves, analyses, critiques, reportages) en ce qui concerne certains aspects de la coréférence, comme par exemple la distance entre anaphore et antécédent, le nombre d'éléments dans une chaîne, etc. Il n'existe à notre connaissance aucun travail qui ait précisément décrit la variation de ces aspects en fonction du genre textuel, même si plusieurs auteurs reconnaissent son existence. Constituer un corpus représentatif de tous les genres textuels du journal Le Monde peut donc s'avérer assez coûteux ; c'est pourquoi nous avons choisi, afin d'alléger cette tâche, de constituer un corpus qui soit homogène du point de vue du genre, composé de textes d'actualités (reportages et comptes-rendus).

Le genre de texte choisi est, quant à lui, motivé par deux facteurs. Premièrement, étant donné les utilisations possibles de notre travail chez France Télécom, notamment pour l'extraction d'informations et le résumé automatique, il nous a semblé judicieux de nous concentrer sur les textes les plus susceptibles de faire l'objet d'applications de ce type. En second lieu, les textes d'actualités sont souvent soumis à certaines contraintes stylistiques ou éditoriales, notamment en ce qui concerne la complexité des phrases ; on peut donc supposer que ces contraintes ont des effets mesurables quantitativement. Ainsi, pour chacun des 8 561 textes du corpus du journal *Le Monde*, nous avons calculé deux valeurs : le nombre de mots et la moyenne de mots par phrase. Ensuite la moyenne et l'écart-type ont été calculés pour l'ensemble du corpus. Les textes comprennent donc en moyenne 474 mots (écart-type : 443), et la moyenne de mots par phrase est de 27 (écart-type : 9). Les 79 textes sélectionnés ont été choisis parmi ceux pour lesquels ces deux valeurs sont situées dans une intervalle entre la moyenne, et la moyenne plus l'écart-type (*i.e.* de 474 à 917 mots).

3. ANNOTATION

L'objectif de l'annotation est de disposer d'un corpus contenant des informations sémantiques, syntaxiques et coréférentielles de base sur les expressions nominales d'un ensemble de textes, ceci afin de pouvoir valider la manière dont sont traitées les informations linguistiques dans les divers algorithmes de résolution existants ; nous ne cherchons pas, à cette étape du travail, à disposer d'un corpus suffisamment riche pour pouvoir être présenté à l'entrée d'un système de résolution.

L'annotation manuelle de la coréférence n'est pas, malgré toutes les précautions qu'on peut prendre, une méthode infaillible pour disposer de résultats fiables : Poesio et Vieira (1998) montrent que de manière générale le choix des types d'objets à annoter peut avoir un impact sur l'accord entre annotateurs. Plus spécifiquement, dans le contexte de leurs expériences, ils constatent que le désaccord entre annotateurs en ce qui concerne la détection d'un antécédent pour une expression donnée provient principalement de ce que cette dernière expression peut être perçue à la fois comme une reprise et comme un cas de « bridging » : l'appartenance d'une expression à plus d'une catégorie semble ainsi être à l'origine du désaccord. Ce constat trouve un écho dans les travaux sur l'annotation sémantique de corpus : certains auteurs (Kilgariff, 1998 ; Veronis 1998) montrent en effet que la polysémie est une source de désaccord entre annotateurs. Il s'agit donc, pour minimiser ces divergences, de sélectionner un ensemble de classes qui soient mutuellement exclusives. Notons par ailleurs que le problème du désaccord entre annotateurs n'est pas évacué lorsque les textes à annoter le sont par une seule personne : il peut exister des différences entre deux annotations d'un même texte faites par une même personne à deux moments distincts.

Chaque texte à annoter est lemmatisé et mis sous format XML. L'annotation consiste alors à ajouter manuellement une balise <EXPRESSION> autour des terminaux qui constituent une expression référentielle. Cette balise a les attributs NPR, TYP, STX et IDREF, qui désignent respectivement le statut de nom propre, la classe sémantique, la catégorie syntaxique et un identifiant associés à l'expression référentielle. Les valeurs possibles de TYP (spécifiées dans le tableau 1) sont inspirées de Bauer (1985, cité par Grass, 2001).

PERS	les noms de personnes
LIEU	les entités géographiques, topologiques, ou de bâtiments
ORG	les regroupements humains (entreprises, associations, etc.)
EV	les évènements de source humaine (évènements sportifs, guerres, etc.)
PROD	les produits industriels ou culturels (*Ford Focus*, *Farenheit 911*, etc.)
PHEN	les phénomènes naturels ou de société (*La comète de Halley*, *l'ouragan Mitch*, *le Bouddhisme*, etc.)
TEMP	les désignations temporelles (*le mercredi 2 octobre*, *en 2003*, *le mois dernier*)
EXPL	les pronoms *il* et *on* sans référent
NONE	les noms sans référent concret (une méthode de *travail*, la résidence du *président*)

Tableau 1 – Classes sémantiques

Malgré le fait que les classes sémantiques soient *a priori* mutuellement exclusives, il est assez fréquent de se retrouver devant des cas de métonymie caractérisés par une ambiguïté entre ORG et LIEU (« L'Aéroport JFK embauche 135 000 personnes et occupe 20 km ») ou entre ORG et EV (« Le Festival d'Angoulême reçoit plus de 500 artistes et dure une semaine »). Afin d'éviter les problèmes de désaccord, on se doit de traiter ces cas de manière uniforme, *i.e.* éviter de laisser à l'annotateur le choix de l'interprétation. Ainsi lorsque deux lectures sont possibles nous avons choisi de favoriser systématiquement le type le moins représenté, soit ORG dans les ambiguïtés LIEU/ORG et EV pour ORG/EV. Ce faisant nous atténuons les différences de distribution entre types sémantiques pour chaque catégorie syntaxique. Nous préférons en effet éviter de les accroître artificiellement, et nous croyons par ailleurs que, malgré tout, les différences intéressantes entre ces distributions seront maintenues.

Les valeurs que l'on donne à l'attribut STX, décrites dans le tableau 2, correspondent quant à elles aux classes syntaxiques possibles des expressions référentielles, et contiennent par ailleurs des informations sur leur contexte lorsque celles-ci correspondent à des syntagmes nominaux pleins.

PPER	Pronom personnel sujet
PCLI	Pronom personnel objet (Les prédictions *le* donnent vainqueur)
PREL	Pronom relatif
PDEM	Pronom démonstratif (*Celui-ci*)
PPOS	Pronom possessif (*Le sien*)
PIND	Pronom indéfini (L'*un* préfère rester, l'*autre* non)
DDEF	GN défini (Le chef d'entreprise)
DIND	GN indéfini (Un chef d'entreprise)
DDEM	GN démonstratif (Ce chef d'entreprise)
DPOS	GN possessif (Ses actionnaires)
SDPR	GN sans déterminant précédé d'une préposition (Projet de *société*)
SDVB	GN sans déterminant précédé d'un verbe ([...] est élu *député*)
SDAP	GN sans déterminant en apposition (Le sénateur Dole, *leader* républicain)
SDXX	Autres GN sans déterminant

Tableau 2 – Classes syntaxiques

Nous avons ainsi étiqueté 14 486 expressions nominales et 6 512 chaînes de coréférence. Environ 46 % de toutes les expressions nominales font partie d'une chaîne, et

un peu plus de la moitié (53 %) de ces chaînes réfèrent à des entités nommées, c'est-à-dire qu'elles sont désignées par un nom propre au moins une fois dans le texte.

4. QUELQUES RÉSULTATS

Les premières données extraites concernent la distribution des classes sémantiques dans l'ensemble des expressions en fonction de leur classe syntaxique. Dans le Graphique 1 chaque colonne désigne une classe syntaxique où la proportion d'expressions de chaque classe sémantique est représentée par une couleur différente – pour des raisons de clarté, nous n'avons distingué que les trois classes sémantiques les plus fréquentes. On remarque d'abord que les pronoms personnels sont plus fréquemment utilisés, dans l'absolu, pour désigner les personnes (38 % des cas) que les organisations (5 %) ou les lieux (1 %). Dans l'hypothèse qu'il n'existerait aucune corrélation entre les classes sémantique et syntaxique, la distribution attendue correspondrait à celle des classes sémantiques toutes classes syntaxiques confondues, soit : PERS : 13 %, ORG : 7 %, LIEU : 8 %, AUTR : 66 %. Il existe donc une certaine affinité entre PPER et PERS ; on remarque des affinités semblables entre les paires de catégories PERS / SDAP, LIEU / SDPR et ORG / DDEF.

Graphique 1 – Distribution des catégories syntaxiques par classe sémantique

Le deuxième ensemble de données extraites, dans le Graphique 2, montre la variation de la distance entre deux expressions en fonction de la classe syntaxique de la deuxième - pour une expression référentielle donnée de classe syntaxique connue, nous voulions connaître la taille de la fenêtre de recherche dans laquelle il est le plus probable de trouver son antécédent. Nous avons conservé quatre classes syntaxiques assez distinctes à cet égard, encore une fois à des fins de clarté ; chaque courbe dans le graphique correspond à une des classes. Les intervalles de distance sont de cinq mots, représentés sur l'axe horizontal; l'axe des ordonnées correspond au nombre d'occurrences dans chaque intervalle de distance. Par exemple pour la courbe correspondant aux pronoms relatifs (PREL), on constate que plus de 80 % d'entre eux (qui font partie d'une chaîne de coréférence)

apparaissent entre 1 et 5 mots à la suite de leur antécédent; un peu plus de 10 % sont situés entre 6 et 10 mots plus loin ; etc. On remarque donc que la distance entre deux expressions est un facteur à prendre en compte lors de la résolution de la coréférence ; ce constat n'a rien de révolutionnaire, cependant on note une grande variation dans l'importance que peut avoir ce facteur selon la classe syntaxique de la reprise, jusqu'à être sans effet dans le cas des syntagmes nominaux définis.

Graphique 2 – Distance en fonction des propriétés catégorielles de la reprise

Au-delà de ces dépendances spécifiques entre propriétés des expressions coréférentes, les données extraites semblent montrer de manière plus générale que les informations dont on se sert pour alimenter les algorithmes de résolution peuvent difficilement être considérées en isolation les unes des autres. Ceci jette un doute sur la validité de certaines approches (Hartrumpf, 2001 ; Popescu-Belis, 1999 ; Cardie et Wagstaff, 1999) où la relation de coréférence est établie de manière exclusivement symbolique, que ce soit par application de règles, vérification de critères ou comparaison des propriétés des expressions en jeu. En effet, s'il est vrai qu'un grand nombre de propriétés des expressions référentielles sont ainsi interdépendantes, les méthodes symboliques devront se baser sur une description (pour chaque langue à traiter) de chacune de ces dépendances, ce qui rend la tâche considérablement plus lourde. En revanche, plusieurs exemples de méthodes statistiques, notamment celles qui sont fondées sur des arbres de décision ou des algorithmes probabilistes (*e.g.* Soon *et al*., 2001 ; Ge *et al*., 1998) permettent d'éviter cette description : les corrélations entre variables sont en effet apprises par l'algorithme lors d'une phase d'entraînement.

4. CONCLUSION

Nous avons présenté ici les premières étapes d'un travail visant le développement d'un système de résolution de la coréférence. Nous avons décrit le corpus mis sur pied à cette fin, puis le schème d'annotation de la coréférence qui lui a été appliqué. De cette annotation nous avons extrait quelques données qui semblent indiquer que l'approche

statistique est à préconiser pour le problème de la résolution de la coréférence, plutôt qu'une approche uniquement symbolique.

5. RÉFÉRENCES

Bauer G. 1985. *Namekunde des Deutschen.* Bern. Germanistische Lehrbuchsammlung Band 21.

Cardie C. et Wagstaff K. 1999. 'Noun phrase coreference as clustering'. *Joint SIGDAT conference on empirical methods in natural language processing and very large corpora (EMNLP-VLC99).*

Ge, N., Hale, J. et Charniak, E. 1998 'A Statistical Approach to Anaphora Resolution'. *Proceedings of the Fifth Workshop on Very Large Corpora.* Hong-Kong, 1998.

Grass T. 2000. 'Typologie et traductabilité des noms propres de l'allemand vers le français'. *Traitement Automatique des Langues* 41(3).

Hartrumpf, S. 2001. 'Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics'. *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001).*

Huang Y. 2000. *Anaphora : A Cross-linguistic Study*. Oxford Studies in Typology and Linguistic Theory. Oxford University Press.

Kilgariff, A. 1998. 'Inter-tagger Agreement'. *Programme and advanced papers of the Senseval workshop.*

Lappin S. et Leass H. 1994. 'An Algorithm for Pronominal Anaphora Resolution'. *Computational Linguistics* 20(4).

Poesio M. et Vieira R. 1998. 'A Corpus-based Investigation of Definite Description Use'. *Computational Linguistics* 24(2).

Popescu-Belis, A. 1999. *Modélisation multi-agent des échanges langagiers : application au problème de la référence et à son évaluation.* Thèse, Université Paris-XI.

Soon, W.M., Ng, H., Lim, D.C.Y. 'A Machine Learning Approach to Coreference Resolution of Noun Phrases'. *Computational Linguistics* 27(4).

Véronis, J. 1998. 'A Study of Polysemy Judgments and Inter-annotator Agreement'. *Programme and advanced papers of the Senseval workshop.*

DEUX INDICES POUR L'ÉTUDE DES MARQUEURS DE LA CONTINUITE THÉMATIQUE DANS DE GRANDS CORPUS

Sophie Piérard et Yves Bestgen[1]
Université catholique de Louvain / Fonds national de la recherche scientifique

1. OBJECTIF ET ÉTAT DE LA QUESTION

L'objectif de ce travail est de comparer, par l'analyse de grands corpus de textes, la fonction d'indicateur de la structure du discours de plusieurs types d'éléments linguistiques introduits en début de phrase tels les adverbiaux temporels, les adjectifs possessifs ou démonstratifs. À la base d'une telle étude se trouve l'idée qu'un texte n'est pas qu'une suite linéaire de phrases, mais qu'il se laisse diviser en segments regroupant des ensembles de phrases qu'on dira localement cohérents. Les relations entre les segments renvoient à ce qu'on a coutume d'appeler la cohérence globale. La structure d'un discours est déterminée par ces deux types de cohérence. Ce sont les éléments linguistiques qui signalent cette structure que nous étudions. Une meilleure connaissance des marqueurs de la structure devrait permettre d'améliorer l'intelligibilité de textes puisque l'usage adéquat de ces marqueurs souligne la cohérence d'un texte et en améliore la représentation mentale que s'en fait le lecteur/auditeur (par exemple : Halliday et Hasan, 1976 ; van Dijk et Kintsch, 1983). Ces mêmes marques sont des sources d'information potentiellement très intéressantes pour les systèmes de segmentation automatique des textes comme l'ont montré Passonneau et Litman (1997) et Beeferman *et al.* (1999).

Les éléments linguistiques qui signalent la structure du discours retiennent depuis de nombreuses années l'attention des chercheurs et une longue liste de candidats à cette fonction a été proposée (Bestgen, 1998 ; Longacre, 1979 ; Marcu, 2000 ; Virtanen, 1992). Les arguments empiriques présentés pour les soutenir sont généralement issus de l'analyse d'un nombre très réduit de textes. La raison en est que l'étude de ces marqueurs présuppose la connaissance de la 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. La complexité et le coût de ces procédures manuelles rendent l'étude de grands corpus impraticable. La conséquence de cette situation est que l'on dispose actuellement de peu de données issues de grands corpus qui attestent qu'un élément donné est plus fréquemment employé en situation de rupture ou en situation de continuité thématique. Pour dépasser cette limitation, il est nécessaire de s'appuyer sur des techniques qui permettent de déterminer automatiquement cette structure. Dans la présente recherche, nous proposons deux indices susceptibles de combler partiellement cette lacune. Ceux-ci permettent en effet de déterminer si un élément linguistique a tendance à apparaître plus souvent en situation de continuité ou de discontinuité thématique. Nous les avons employés pour comparer dans un corpus de textes littéraires une série de signaux potentiels de la structure d'un texte.

Dans la suite de ce rapport, nous présentons d'abord les deux indices de continuité /discontinuité thématique que nous proposons. Si le premier est simple, le second s'appuie sur des développements récents en segmentation automatique de textes. Ensuite, nous décrivons les éléments linguistiques dont nous souhaitons évaluer la fonction d'indicateur de la structure. Enfin, nous présentons le corpus sur lequel les analyses ont été effectuées et les résultats de celles-ci.

2. MÉTHODOLOGIE PROPOSÉE

Pour déterminer si une expression a tendance à apparaître plus souvent en situation de continuité ou de discontinuité thématique, nous proposons, en premier lieu, d'employer un indice qui traduit, au moins partiellement, les intentions de l'auteur d'un texte : les changements de paragraphe (ou alinéas). L'auteur d'un texte est en effet censé les introduire pour signaler une discontinuité thématique (Hofmann, 1989 ; Longacre, 1979). Nous proposons donc de déterminer si un élément linguistique apparaît plutôt en début de paragraphes qu'au milieu de ceux-ci en utilisant la technique classique du test du Chi² et le rapport des chances qui lui est associé (Howell, 1998 ; Piérard et Bestgen, 2005a ; Rogati et Yang, 2002). Un marqueur de continuité thématique devrait se trouver plus souvent au sein d'un paragraphe alors qu'un marqueur de discontinuité devrait apparaître plus fréquemment en début de paragraphe.

L'inconvénient majeur de ce premier indice est que les paragraphes remplissent d'autres fonctions discursives (Stark, 1988 ; Brown et Yule, 1983), comme par exemple, la mise en évidence d'un élément du texte. Nous proposons donc d'employer parallèlement un indice basé sur la cohésion lexicale qui a, entre autres, été employé avec succès pour segmenter automatiquement des textes (Bestgen, 2005, sous presse ; Choi *et al.*, 2001).

Ce second indice est issu de l'analyse sémantique latente, une technique mathématique qui vise à extraire un espace sémantique de très grande dimension à partir de l'analyse statistique de l'ensemble des cooccurrences dans un corpus de textes (Deerwester *et al.*, 1990 ; Landauer et Dumais, 1997). Comme le souligne Landauer *et al.* (1998), cette technique peut être vue de deux manières. À un niveau théorique, elle peut servir de base pour développer des simulations des processus psycholinguistiques à l'œuvre lors de la compréhension du langage (Landauer *et al.*, 1998), incluant, par exemple, un « modèle computationnel » du traitement des métaphores (Kintsch, 2000). À un niveau plus appliqué, c'est une technique permettant d'inférer et de représenter le sens de mots sur la base de leur usage dans des textes, mais aussi d'analyser la cohérence dans des textes (Foltz *et al.*, 1998 ; Bestgen, Degand et Spooren, sous presse).

Le point de départ de l'analyse est un tableau lexical qui contient le nombre d'occurrences de chaque mot dans chaque segment de textes, un segment pouvant être un paragraphe ou une phrase ou même une suite de mots d'une longueur arbitraire. Ce tableau fait l'objet d'une décomposition en valeurs singulières, une sorte d'analyse factorielle, qui en extrait les dimensions orthogonales les plus importantes. Les milliers de mots caractérisant les documents sont ainsi remplacés par des combinaisons linéaires ou « dimensions sémantiques » sur lesquelles peuvent être situés les mots originaux. Contrairement à une analyse factorielle classique, les dimensions extraites sont très nombreuses (plusieurs centaines) et non interprétables. Elles peuvent toutefois être vues comme analogues aux traits sémantiques fréquemment postulés pour décrire le sens des mots (Landauer et al., 1998). Dans cet espace, le sens de chaque mot (phrases ou

paragraphe) est représenté par un vecteur. Pour calculer la similarité sémantique entre deux mots (phrases ou paragraphes), on calcule le cosinus entre les vecteurs qui les représentent. Plus deux phrases sont sémantiquement proches, plus leur cosinus est élevé, la valeur maximale étant 1. Un cosinus de 0 indique une absence de similarité. L'intérêt majeur de cet espace sémantique est qu'il permet d'identifier comme similaires deux passages d'un texte même si ceux-ci ont peu de mots en commun.

Nous avons employé cette technique pour déterminer si une expression apparaît plus souvent en situation de continuité ou de discontinuité. Plus précisément, nous nous sommes basés sur les cosinus entre la phrase qui commence par cette expression, appelée ici phrase cible (p), et deux autres phrases : celle qui la précède (p-1) et celle qui la suit (p+1). Plus le cosinus entre la phrase cible et celle qui la précède (cos[p-1,p]) est grand, plus la phrase cible est en situation de continuité thématique. Si une phrase cible introduit un changement de thème, c'est le cosinus avec la phrase qui la suit (cos[p, p+1]) qui devrait être plus élevé. En comparant ces deux cosinus, il est donc possible d'évaluer la fonction d'indicateur de la structure d'une expression.

3. ÉLÉMENTS LINGUISTIQUES ANALYSÉS

L'objectif premier de ce travail étant de tester l'efficacité des indices proposés, nous avons choisi d'analyser des éléments linguistiques qui, selon la littérature, devraient fonctionner comme des signaux de la structure d'un texte. Il s'agit tant d'éléments censés signaler des ruptures thématiques comme les adverbiaux temporels cadratifs (« le matin ») ou les groupes nominaux introduits par un article indéfini (« un … », « des … ») que d'éléments censés signaler une continuité thématique tels les pronoms personnels (« il », « elles »), les adjectifs possessifs (« son », « ses », « sa ») et démonstratifs (« cet », « ces ») et les groupes nominaux introduits par un article défini (« le… », « l'… »). Linguistiquement, ces marques de continuité et de rupture sont unies par leur propriété référentielle très largement non autonome. Étudier simultanément ces deux types d'éléments est une des originalités de ce travail. Jusqu'à présent en effet, les recherches, effectuées sur de petits corpus, ont principalement porté sur des dispositifs marquant une rupture comme les expressions adverbiales temporelles et spatiales (Charolles, 1997 ; Virtanen, 1992). Si les marques qui signalent la continuité discursive ont été largement étudiées en linguistique, peu de travaux ont évalué leur fonction de signaux du discours et aucun, à notre connaissance, ne s'est intéressé à l'expression « adjectif possessif + nom ». La spécificité de cette dernière est qu'elle est biréférentielle : elle représente un référent par rapport à un autre (Apothéloz, 1995 ; Broeder, 1992). Bien que le nom soit l'élément principal, l'adjectif possessif devrait souligner la continuité avec les phrases précédentes (Piérard et Bestgen, 2005b). Comparer cette expression avec les adjectifs démonstratifs (Botley et McEnery, 2001) et les pronoms personnels (Delisle, 1993) sera particulièrement instructif.

4. CORPUS

Le corpus est composé de textes littéraires extraits des bases ABU, Intratext et Wordthèque. Il contient 67 romans (du XIXe et XXe siècle) et approximativement 4.300.000 mots. Les textes ont été découpés en phrases et lemmatisés au moyen du programme *TreeTagger* de Schmid (1994). Comme on peut penser que les marqueurs de la structure sont différents à l'oral et à l'écrit, nous avons retiré du corpus les paragraphes qui

contenaient des dialogues afin de focaliser les analyses sur l'emploi des indicateurs de la structure à l'écrit.

Le tableau 1 donne la liste des éléments linguistiques étudiés et le nombre de ceux-ci apparaissant au tout début d'une phrase. Comme adverbiaux cadratifs, nous avons employé les expressions temporelles incluant une référence à une des grandes subdivisions de la journée (*matin, matinée, après-midi, soir, soirée, nuit*) dont Costermans et Bestgen (1991) ont montré la fonction de marqueurs de discontinuité thématique. En ce qui concerne les pronoms personnels et les adjectifs possessifs, nous n'avons pris en compte que la troisième personne.

	Fréquence
Adverbiaux cadratifs temporels (CAD)	581
Articles définis (DEF)	10343
Articles indéfinis (IND)	2167
Adjectifs démonstratifs (DEM)	1792
Adjectifs possessifs 3e pers. (POS)	1478
Pronoms personnels 3e pers. (PPE).	12915

Tableau 1 : Fréquence dans le corpus des expressions linguistiques analysées

Ce même corpus, dans version lemmatisée et expurgée des dialogues, a été employé pour construire l'espace sémantique nécessaire pour l'analyse sémantique latente. Spécifiquement pour cette analyse, les formes fonctionnelles (pronoms, articles, ...) ont été supprimées. Ce corpus a été segmenté en unités de 60 à 120 mots composées de phrases entières. Tous les mots dont la fréquence dans le corpus était au moins égale à 2 ont été conservés. La matrice de cooccurrences ainsi obtenue a été décomposée en valeurs singulières par le programme Svdpack (Berry, 1992) et les 300 premiers vecteurs propres ont été conservés.

5. RESULTATS

5.1 Analyse de la position dans les paragraphes

Pour analyser nos données, nous avons choisi d'utiliser la technique du rapport des chances. Le rapport des chances présente l'avantage de permettre les comparaisons entre expressions parce que, contrairement au Chi², il n'est pas affecté par des totaux inégaux de ligne et donc par le fait que certaines expressions sont nettement plus fréquentes que d'autres (Howell, 1998, p. 182). Il s'agit, dans le cas qui nous occupe, du rapport entre la chance qu'une expression arrive en tête de paragraphe par rapport à la chance que tout autre mot arrive en tête de paragraphe. Ce rapport des chances (RC) est calculé à partir d'une table de contingence, comme présentée dans le tableau 2. Cet exemple, basé sur des données recueillies lors de la présente étude, une phrase qui commence par un adjectif possessif a 2 fois moins de chances d'être en tête de paragraphe qu'au milieu :

$$RC = \frac{297/1684}{27704/77997} = 0.50$$

	Début de paragraphe	Pas début de paragraphe	
Possessif	297	1 684	1 981
Autres expressions	27 704	77 997	105 701
	28 001	79 681	107 682

Tableau 2 : table de contingence pour les adjectifs possessifs de la 3ème personne

Un rapport des chances inférieur à 1 indique donc qu'une expression apparaît plus souvent en milieu de paragraphe qu'en début. Un rapport des chances supérieur à 1 indique l'inverse : l'expression apparaît plus souvent en début de paragraphe qu'en milieu. Plus le rapport des chances est différent de 1 et plus cette différence dans la distribution de l'expression en fonction de sa position est importante. Afin de déterminer à partir de quelle valeur un rapport des chances est suffisamment éloigné de 1 pour pouvoir être considéré comme significatif, nous avons employé le classique test du Chi².

La figure 1 présente l'ensemble des résultats obtenus avec la mesure du Chi² et du rapport des chances. L'axe gradué indique dans quelle proportion une expression est observée plus souvent en début de paragraphe qu'au milieu (côté droit - rapport supérieur à 1) ou l'inverse (côté gauche - rapport inférieur à 1). Par exemple, la valeur « 3/1 » indique qu'un élément à trois fois plus de chance d'apparaître en début de paragraphe qu'au milieu ; la valeur « 1/2 » indique qu'un élément a deux fois plus de chance d'apparaître au milieu d'un paragraphe qu'au début, la valeur « 1/1 » signalant un élément apparaissant aussi souvent au début qu'au milieu. Les tests du Chi² indiquent que tous les éléments linguistiques analysés ont un rapport des chances significativement différent de 1/1. Selon cet indice, ils sont donc tous des indicateurs de continuité ou de rupture.

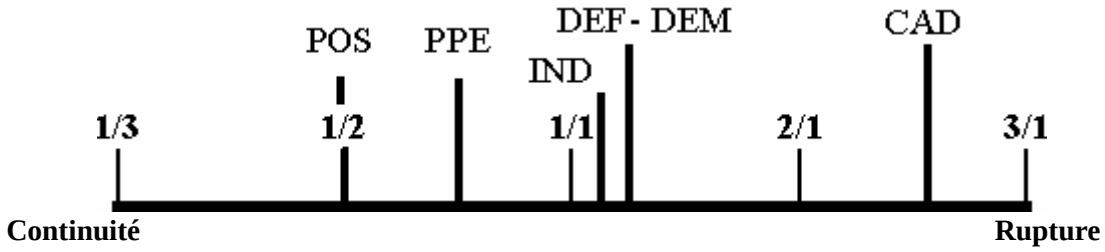


Figure 1 : distribution des expressions en fonction du rapport des chances.

Globalement, ces résultats correspondent à nos attentes et apportent donc une première validation de la technique. Les adverbiaux cadratifs temporels sont bien des marques de discontinuité et les pronoms personnels des marques de continuité. L'article indéfini est bien du côté discontinu de l'échelle. On note toutefois qu'il est très proche du point neutre, même s'il en est significativement différent selon le test du Chi². On note surtout que

l'article défini et l'adjectif démonstratif s'observent plus fréquemment que lui en début de paragraphe. Ce résultat, moins attendu, sera discuté après l'analyse du second indice. Enfin, l'adjectif possessif fonctionne comme les pronoms.

5.2 Analyse de la proximité thématique

La figure 2 présente les résultats obtenus par l'analyse de la proximité thématique. L'axe gradué indique la valeur du cosinus. Plus cette valeur est élevée, plus les phrases analysées sont thématiquement proches. La partie supérieure de la figure présente les cosinus moyens entre la phrase qui commence par l'élément en question et la phrase qui la précède, la partie inférieure rapportant les cosinus moyens entre cette même phrase cible et celle qui la suit.

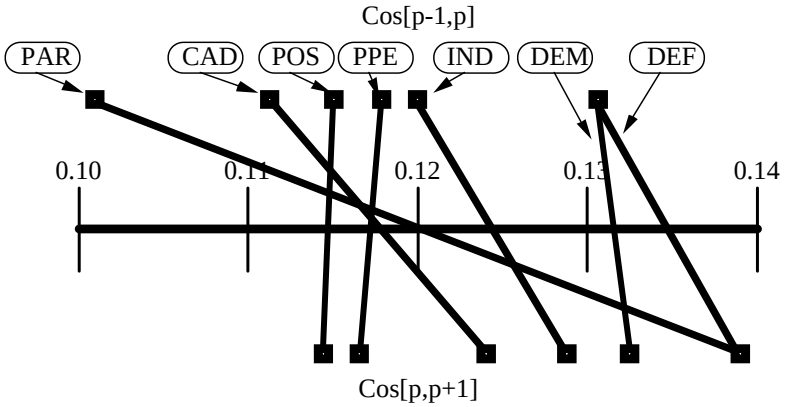


Figure 2 : distribution des expressions en fonction des cosinus.

À titre d'exemple, mais aussi afin de valider la procédure, nous avons calculé ces cosinus moyens pour la première phrase d'un paragraphe (PAR). Comme attendu, on observe un cosinus moyen plus faible (0.10) entre la première phrase d'un paragraphe et celle qui la précède qu'entre cette même phrase et celle qui la suit (0.14), confirmant la fonction de marqueur de rupture du changement de paragraphe.

Les adverbiaux cadratifs et les articles indéfinis présentent également le profil attendu de marques de discontinuité. Les pronoms personnels et les adjectifs possessifs se comportent plutôt comme des marques de continuité, confirmant les résultats de la première analyse.

Les adjectifs démonstratifs et les articles définis méritent une attention particulière. Ils présentent globalement des cosinus très élevés tant avec la phrase qui précède qu'avec celle qui suit. Cette observation permet, selon nous, de mieux comprendre les résultats obtenus lors de l'analyse des changements de paragraphe. On peut en effet penser qu'un certain nombre de changements de paragraphe n’ont pas comme objectif de signaler une rupture thématique, mais plutôt de mettre en évidence le nom qu'ils déterminent (Stark, 1988). Une autre interprétation est que ces éléments sont employés pour souligner la continuité entre deux paragraphes successifs et tout particulièrement le fait que ceux-ci portent sur un thème commun (Hofmann, 1989). Le cas des articles définis est encore plus particulier. Ceux-ci sont encore plus liés avec la phrase qui suit qu'avec celle qui précède, à la manière d'une marque de discontinuité, mais leur cosinus très élevé avec la phrase qui précède rend une telle interprétation difficile. Plus généralement, le fait que les phrases commençant par un article défini, un adjectif démonstratif ou même un article indéfini donnent lieu à des

cosinus moyens plus élevés que ceux observés pour les pronoms personnels et les adjectifs possessifs mérite des investigations complémentaires comme une analyse comparative fine d'échantillons de ces expressions.

6. CONCLUSION

Nous avons présenté deux indices qui permettent d'analyser dans de grands corpus la fonction d'indicateur de la structure d'éléments linguistiques : un indice basé sur les changements de paragraphe et un indice de cohésion lexicale. Les résultats obtenus grâce à ces deux indices sont globalement en accord avec la littérature : les éléments linguistiques censés indiquer un changement de thème sont bien repérés comme tels. Il en est de même pour les expressions signalant des continuités. Cette première validation autorise une analyse plus fine des résultats et, plus particulièrement, la comparaison des différentes expressions entre elles. Ainsi, le déterminant possessif apparaît moins souvent en début de paragraphe que le pronom personnel, soulignant sa fonction de signal de continuité, en accord avec les analyses linguistiques montrant que l'antécédent du déterminant est presque toujours au plus loin dans la phrase qui précède (Apothéloz, 1995). La complémentarité des deux procédures est mise en évidence par l'interprétation du démonstratif et du déterminant défini comme des dispositifs signalant un glissement thématique : l'introduction d'un nouveau thème lié au précédent.

Parmi les développements que nous nous proposons d'effectuer, il y a tout d'abord la nécessité de remédier à une limitation de l'indice de cohésion lexicale dont la portée est très locale. Il indique seulement qu'une phrase est sémantiquement plus liée avec celle qui la suit qu'avec celle qui la précède ou l'inverse. Le paragraphe est par contre une mesure plus globale puisqu'il segmente le texte en paquets de phrases qui forment à chaque fois une unité textuelle. Nous proposons d'améliorer cet indice en employant les cosinus pour dériver une mesure plus globale de la structure d'un texte basée sur l'algorithme de segmentation automatique de Choi *et al.* (2001 ; Bestgen, sous presse). Une autre limite de cette recherche qui appelle des développements réside dans l'absence d'analyse linguistique fine des éléments considérés. Il serait aussi utile d'effectuer les mêmes analyses sur des corpus d'autres genres comme des articles de journaux afin d'évaluer la généralité des résultats obtenus.

7. REFERENCES

Apothéloz D. 1995. *Rôle et fonctionnement de l'anaphore dans la dynamique textuelle*. Genève : Droz.

Beeferman D. Berger A. & Lafferty J. 1999. 'Statistical models for text segmentation'. *Machine Learning*, 34, p. 177–210.

Berry M.W. 1992. 'Large scale singular value computation'. *International journal of Supercomputer Application*, 6, p. 13-49.

Bestgen Y. 1998. 'Segmentation markers as trace and signal of discourse structure'. *Journal of Pragmatics*, 29, p. 753-763.

Bestgen Y. 2005. 'Amélioration de la segmentation automatique des textes grâce aux connaissances acquises par l'analyse sémantique latente' in Actes *de la conférence TALN 2005*, Dourdan, juin 2005.

Bestgen Y. (sous presse). *Improving Text Segmentation Using Latent Semantic Analysis : A Reanalysis of Choi, Wiemer-Hastings and Moore (2001). Computational Linguistics*.

Bestgen Y., Degand L., & Spooren W. (sous presse). *Towards automatic determination of the semantics of connectives in large newspaper corpora. Discourse Processes*.

Botley S. & McEnery T. 2001. 'Demonstratives in English : a corpus-based study'. *Journal of English Linguistics*, 29, p. 7-33.

Broeder P. 1992. 'Possession in a new language'. *Applied Linguistics*, 13 (1), p. 100-118.

Brown G. & Yule G. 1983. *Discourse analysis*. Cambridge : Cambridge University Press.

Charolles M. 1997. 'L'encadrement du discours - univers, champs, domaines et espaces'. *Cahier de recherche linguistique*, 6, p. 1-73.

Choi F., Wiemer-Hastings P. & Moore J. 2001. 'Latent semantic analysis for text segmentation'. in *Proceedings of NAACL'01*. p. 109–117.

Costermans J. & Bestgen Y. 1991. 'The role of temporal markers in the segmentation of narrative discourse'. *CPC: European Bulletin of Cognitive Psychology*, 11, p. 349-370.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. & Harshman R. 1990. 'Indexing by Latent Semantic Analysis'. *Journal of the American Society for Information Science*, 41, p. 391-407.

Delisle H. 1993. 'Anaphora in German Discourse : an empirical study'. in Eid M. et Iverson G. (Eds). *Principles and prediction : the analysis of natural language.* Amsterdam : John Benjamins Publishing Compagny.

Foltz P.W., Kintsch W. & Landauer T.K. 1998. ‚The measurement of textual coherence with Latent Semantic Analysis'. *Discourses Processes*, 25, p. 285-307.

Halliday M.A.K. & Hasan R. 1976. *Cohesion in English*. Landon : Longman.

Hofmann T.R. 1989. 'Paragraphs, & anaphora'. *Journal of Pragmatics*, 13, p. 239-250

Howell D.C. 1998. *Méthodes statistiques en sciences humaines*. Bruxelles : De Boeck Université.

Kintsch W. 2000. 'Metaphor comprehension : a computational theory'. *Psychonomic Bulletin and Review*, 7, p. 257-266.

Landauer T.K. & Dumais S.T. 1997. ‚A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge'. *Psychological Review*, 104, p. 211-240.

Landauer T.K., Foltz P.W. & Laham D. 1998. 'An introduction to Latent Semantic Analysis'. *Discourse Processes*, Vol. 25, p. 259-284.

Longacre R. E. 1979. 'The paragraph as a grammatical unit'. in Givón T. (ed.), *Syntax and Semantics, 12 : Discourse and Syntax*, New York : Academic Press, p. 115-134.

Marcu D. 2000. 'The rhetorical parsing of unrestricted texts : A surface-based approach'. *Computational Linguistics*, 26, p. 395-448.

Passonneau R.J. & Litman D.J. 1997. 'Discourse Segmentation by Human and Automated Means'. *Computational Linguistics*, 23, p. 103-139.

Piérard S. & Bestgen Y. 2005a. 'Identification automatique des marqueurs globaux du discours par l'analyse des expressions récurrentes'. *Phraseology 2005*, Louvain-la-Neuve, octobre 2005.

Piérard S. & Bestgen Y. 2005b. 'When production meets comprehension : effects of the possessive anaphora and thematic break on a story continuation task'. in *Annual Meeting of the Society for Text and Discourse*, Amsterdam, juillet 2005.

Rogati M. & Yang Y. 2002. 'High-Performing Feature Selection for Text Classification'. *CIKM'02*, November 4–9, McLean, Virginia, USA.

Schmid H. 1994. 'Probabilistic Part-of-speech tagging using decision trees'. in *Proceedings of International Conference on New Methods in Language Processing*.

Stark H.A. 1988. 'What do paragraph markings do ?'. *Discourse Processes*, 11, p. 275-303.

van Dijk T. & Kintsch W. 1983. *Strategies of discourse comprehension*. New York : Academic.

Virtanen T. 1992. *Discourse functions of adverbial placement in English*. Åbo : Åbo Akademi University Press.

8. NOTES

1. Yves Bestgen est chercheur qualifié du Fonds national de la recherche scientifique (FNRS). Cette recherche est financée par une "Action de Recherche concertée" du Gouvernement de la Communauté française de Belgique et par le projet FRFC n° 2.4535.02.

PRINCIPE DE COMPOSITION SEMANTICO-DISCURSIF DES COLLOCATIONS LEXICALES : DE L'ARBITRAIRE AU MOTIVÉ ?

Estelle Dubreil
Laboratoire Informatique de Nantes Atlantique – LINA FRE CNRS 2729

1. PROBLÉMATIQUE

Traditionnellement, l'association terminologique préexistant au phénomène collocationnel est qualifiée d'arbitraire (Grossmann & Tutin, 2003) et rapportée de ce fait à une question d'usage linguistique conventionnel. Pourtant, en soumettant les collocations à une analyse sémantico-discursive du type SPA – Sémantique des Possibles Argumentatifs – (Galatanu, 2004), il s'avère que cette association terminologique est parfois légitime, prévisible puisque sémantiquement *calculable*, ce qui remet en cause leur caractère arbitraire.

Nous avons constitué un corpus électronique spécialisé, le corpus TAL(N), représentatif de la communauté du Traitement Automatique du Langage Naturel (Dubreil, 2004). Ce domaine de pratique scientifique est dynamique en France. Les communications sont nombreuses, essentiellement écrites, ce qui satisfait nos besoins en analyses textuelles et offre une double perspective scientifique confortable. D'une part, le regroupement d'une quantité de données suffisante est possible, et d'autre part, cela laisse espérer une certaine diversité des auteurs, permettant d'accéder à un panel lexical varié. L'exploitation légale des documents sous la forme numérisée n'a pas posée de problème ; les textes appartiennent à l'association loi 1901 du domaine, l'*ATALA*, et ont été délivrés sur autorisation. Le corpus TAL(N), comporte environ 1 million d'occurrences sans la ponctuation. Il réunit 193 articles de recherche scientifique : 69 articles de revue *TAL* et 124 articles de conférence. Il est constitué de textes entiers exclusivement rédigés en français et relevant d'un même registre dit *écrits scientifiques*. Tous les textes appartiennent à une même période consécutive d'une durée égale : la revue *TAL* 1998-2002, du Vol. 39.1 au Vol. 43.3 inclus et la conférence TALN, les actes de 1999 à 2003 inclus.

L'aspect informatisé de notre corpus et le recours nécessaire aux techniques d'analyses liées à la linguistique de corpus, nous ont conduit à observer les phénomènes collocationnels. Dès lors, il nous a paru pertinent de rapprocher la collocation d'un autre concept fondateur en SPA, le possible argumentatif. Ce rapprochement a été motivé par l'intuition forte qu'une collocation correspond potentiellement à la concrétisation d'un possible argumentatif dans le discours.

Après avoir défini ce que nous entendons par collocation et présenté le cadre théorique sémantico-discursif retenu pour cette étude, nous expliciterons l'intuition sous-jacente au rapprochement des deux concepts mobilisés. Puis, sur la base de nos hypothèses de travail et de notre méthodologie, nous détaillerons l'analyse des collocations de notre corpus basées sur *amélioration(s)*. Nous conclurons sur ces résultats.

2. CADRE THÉORIQUE

2.1. Les collocations

2.1.1. LES COLLOCATIONS. Concept général. D'après Williams (2003) :

> « Le terme *collocation* a été introduit dans les années 30 par J. R. Firth, membre fondateur de l'école contextualiste britannique, pour caractériser certains phénomènes linguistiques de cooccurrence qui relèvent essentiellement de la compétence linguistique des locuteurs natifs (Firth, 1957). »

Ce terme recouvre en fait des phénomènes linguistiques très variés relevant de la phonologie, pragmatique, sémantique, syntaxique et même statistique (van der Wouden, 1997). La collocation est la plupart du temps définie comme une association syntagmatique restreinte trouvant sa place entre le syntagme libre et le syntagme figé. Par exemple, en Français un *café noir* est un café sans adjonction de lait et si *café* conserve sa signification habituelle, l'adjectif *noir* prend un sens nouveau. D'ailleurs, en Espagnol, un *café noir* se traduira par un *café solo*. Du fait de leur caractère idiomatique et donc du problème qu'elles posent à tous les apprenants d'une langue étrangère, elles ont été décrites et modélisées par de nombreux lexicographes : citons par exemple le *BBI Dictionary of English Word Combinaison* (Benson et *al*., 1986), le *Dictionnaire explicatif et combinatoire* (Mel'čuk et *al*. 1984-1999).

2.1.2. LES COLLOCATIONS LEXICALES. La collocation lexicale telle que définie par Haussmann (1997) est composée d'une base et d'un collocatif ; dans l'exemple ci-dessus, *café* est la base, *noir* est le collocatif. Elle possède un certain nombre de propriétés : récurrence, arbitraire et bonne formation syntaxique, qui ont été exploitées pour leur découverte en corpus (Daille et Williams, 2001). La motivation sémantique des collocations selon Grossmann et Tutin (2003) repose soit sur des procédés sémantiques productifs, comme la métaphore, la comparaison ou la métonymie, soit sur des régularités sémantiques, par exemple *affreux* se voit associé un sens intensif en cooccurrence avec les noms de sentiments de polarité négative.

2.2. Les possibles argumentatifs

2.2.1. LE MODELE SPA. Le Possible Argumentatif (PA) est un concept pivot dans la théorie SPA auquel il donne d'ailleurs son nom. Pour en mesurer la portée, nous exposons brièvement les fondements du modèle. La SPA se donne pour objectif de proposer une description/représentation de la signification lexicale (Galatanu, 2004), basée simultanément sur des informations sémantiques (internes à la langue) et discursives (externes à la langue, donc contextuelles). La signification lexicale est entendue comme :

> « une conceptualisation largement identique d'un individu à l'autre, [formant] une sorte de socle pour une intercompréhension réussie » (Kleiber, 1999).

Elle se présente donc comme un processus d'abstraction concrétisé par une unité lexicale et visant une stabilisation du monde dans la langue. Du point de vue de la SPA, la description de la signification lexicale doit pouvoir rendre compte d'une part, de la *dimension évaluative du langage*, par une représentation des potentiels argumentatifs des unités lexicales et de la portée axiologique associée, et d'autre part, de la *dimension dénotative* du langage général, par la représentation du monde tel qu'il est perçu et modélisé par et dans la langue. Aussi, la SPA décrit la signification lexicale d'un mot comme étant composée de trois strates : un noyau, des stéréotypes et des possibles argumentatifs.

Le *noyau* est une entité sémantique, composée de traits nécessaires de catégorisation (TNC) et de prédicats abstraits. Les TNC rassemblent trois types d'informations : la nature grammaticale du mot (nom, verbe…), ses caractéristiques modales (épistémique, axiologique…) et son orientation axiologique (positive ou négative). Les prédicats abstraits correspondent à des primitifs sémantiques invariants (Putnam, 1994).

Les *stéréotypes* sont également des entités sémantiques. Ils correspondent à la déclinaison des prédicats abstraits du noyau par association avec d'autres invariants (représentations), dans des blocs de signification argumentatifs posés comme naturels (du type : cause-effet, symptôme-phénomènes, but-moyen…). On parle de *blocs d'argumentation interne* à la signification lexicale (Carel & Ducrot, 1999).

2.2.2. LES POSSIBLES ARGUMENTATIFS. Les PA sont des entités discursives. Ils correspondent à l'association *prévisible* de deux unités lexicales liées par une relation d'argumentation et calculable sur la base des enchaînements linguistiques que la signification lexicale est susceptible de générer en contexte. Plus précisément, il s'agit de la concrétisation dans le discours de l'association du mot avec l'un des éléments de ses stéréotypes. On parle de *blocs d'argumentation externe* à la signification lexicale (Carel & Ducrot, 1999), puisque consommés dans le discours. Un bloc d'argumentation externe représente donc l'association dans une structure argumentative, du mot dont on construit la signification lexicale avec l'un des prédicats identifiés comme faisant partie de son argumentation interne. Nous exemplifions ci-dessous les trois strates de la signification lexicale avec le substantif *colère* (Galatanu, 2004).

COLÈRE

TRAITS NECESSAIRES DE CATEGORISATION (TNC)		
TYPE GRAMMATICAL	ZONE MODALE	ORIENTATION AXIOLOGIQUE
Nom – Féminin	Affectif, Moral et Axiologique	Monovalent – Négatif

ENTITÉS SÉMANTIQUES			ENTITÉS DISCURSIVES
NOYAU		STÉRÉOTYPES	POSSIBLES ARGUMENTATIFS
1	Penser à faire mal à soi-même ou à autrui	– Donc penser mal… – Donc mériter punition… – Donc vouloir vengeance…	– Colère intérieure…
2	Eprouver émotion-sentiment négative violent	– Donc état psychique mauvais… – Donc état physique mauvais… – Donc agressivité… – Donc agir violemment…	– Colère folle… – Violente colère…

Tableau 1 : Signification lexicale de colère

De par le lien argumentatif posé, les PA sont par essence évaluatifs. Ils s'organisent autour de deux faisceaux orientés respectivement vers l'un ou l'autre des pôles axiologiques, positif ou négatif. Dans l'exemple *Elle est intelligente, donc douée à l'école*, on constate l'activation d'un PA axiologique positif, porté par la signification lexicale de

l'adjectif *intelligente*. À l'inverse, dans l'exemple *Il est lâche, donc inutile de compter sur lui en cas d'attaque*, on observe l'activation d'un PA axiologique négatif, porté par la signification lexicale de l'adjectif *lâche*.

Les PA sont des entités discursives, ce qui signifie que le discours influe sur les potentiels argumentatifs axiologiques. D'après Galatanu (2005, à paraître) :

> « Le noyau et les stéréotypes forment dans la SPA un dispositif de génération de séquences discursives argumentatives, « les Posssibles Argumentatifs », qui peuvent être activés dans les occurrences discursives, ou renforcés, voire intervertis, par des phénomènes de contamination co- ou contextuelle. »

Dans l'exemple *Il était en colère, donc il a frappé son ami*, le PA est conforme à l'orientation axiologique des stéréotypes du substantif *colère*. À l'inverse, dans l'exemple *Il était en colère, pourtant il a amorcé une réconciliation*, le PA est contraire à cette même orientation axiologique.

3. HYPOTHÈSES

Partant des définitions de *collocation* au sens d'Haussmann (1997), entendue comme toute séquence textuelle constituée de deux lexèmes dont l'un conserve son sens habituel, syntaxiquement correcte et fréquente, et de *PA*, entendu comme l'association prévisible de deux unités lexicales liées par une relation d'argumentation et calculable sur la base des enchaînements linguistiques que la signification lexicale est susceptible de générer en contexte, nous postulons qu'une collocation correspond potentiellement à la concrétisation d'un PA dans le discours. Cette hypothèse se décline sous la forme d'un double paradigme :

-Soit la collocation repérée est la concrétisation d'un PA. Ex : *violente colère*

Dans la description de la signification lexicale de *colère*, on trouve associé au noyau (entre autres), le stéréotype *violent*. Aussi, la collocation *violente colère* (et inversement *colère violente*) est *calculable* sur la base des PA existants. Dans ce cas, on parlera de *collocation argumentative* et nous vérifierons qu'elle est intrinsèquement axiologisée.

-Soit la collocation ne correspond pas à la concrétisation d'un PA. Ex : *colère froide*

Dans ce cas, l'association terminologique n'est pas prévisible, on ne peut calculer –de façon transparente– *froide* dans *colère*. On parlera alors de *collocation inédite*. En l'absence de relation argumentative directe, la collocation inédite ne devrait pas être axiologisée, ce qui contredit la dimension évaluative du langage. Aussi, nous formulons une nouvelle hypothèse concernant l'axiologisation des collocations inédites, en postulant qu'il existe un processus évaluatif externe dû au discours, effectif par relais contextuel.

> Ex : - *Je ne vois pas très bien, dit le professeur McGonagall avec une colère froide, comment vous pouvez espérer vous faire une idée de mes méthodes d'enseignement si vous persistez à m'interrompre sans cesse. En règle générale, je ne permets à personne de parler en même temps que moi. On aurait dit que le professeur Ombrage venait de recevoir une gifle. Elle ne répondit pas un mot mais ajusta son parchemin sur son bloc-notes et se mit à écrire frénétiquement.* (*Harry Potter et l'Ordre du Phénix*, J.K. Rowling, Ed. Gallimard).

La présence dans le contexte de *persistez à m'interrompre sans cesse* et *recevoir une gifle*, nous permettent de conclure que la collocation *colère froide* possède une orientation

axiologique négative. Naturellement, nous n'excluons pas que la base puisse conserver son potentiel axiologique et le propager ainsi à toute la collocation.

4. MÉTHODOLOGIE

Pour éprouver ces hypothèses, nous avons mis au point un processus de validation pluridisciplinaire, mêlant linguistique de corpus et lexicographie.

4.1. Le niveau linguistique de corpus

Le repérage des unités lexicales à extraire s'effectue sur la base de la liste des mots du corpus TAL(N), générée par le logiciel Wordsmith Tools[1]. Nous avons choisi d'analyser des mots axiologisés au niveau de leur TNC, pour leur capacité à former des blocs d'argumentation. En définitive, sur la base de leur fréquence relativement élevée nous avons retenu une vingtaine d'unités lexicales, parmi lesquelles *amélioration*, *connaissance*, *erreur*, *performance*, *problème*, etc. Nous avons extrait par exemple 134 occurrences de *amélioration* au singulier et 40 occurrences du mot au pluriel.

Ensuite, le dépouillement des données se fait à l'aide du concordancier Contextes[2]. Nous avons épuré nos données en laissant de côté :

- selon un critère statistique, tous les candidats peu fréquents. Nous n'avons conservé que les mots apparaissant au moins trois fois, dans trois articles différents et de trois auteurs différents. Ex : *amélioration constante* ; *nette amélioration*
- selon un critère de pertinence, les candidats composés de mots grammaticaux, pour leur incapacité à former des relations d'argumentation. Ex : *amélioration de/du/des*
- selon un critère lexical, toutes les lexies. Nous n'avons pas pris en compte les expressions figées faisant l'objet d'une entrée dans le dictionnaire. Ex : *à ma/notre connaissance*
- selon un critère lexicographique, les termes complexes associés au domaine de pratique ; vérification faite dans deux dictionnaires terminologiques disponibles sur le Web[3]. Ex : *base(s) de connaissance* ; *connaissance du monde* ; *extraction de connaissances*

4.2. Le niveau lexicographique

Au niveau lexicographique, le protocole de construction de la signification lexicale consiste à prendre les définitions d'un mot dans deux dictionnaires différents mais comparables et à ne retenir pour la construction du noyau et des stéréotypes que ce qui est commun aux différentes entrées. Nous utilisons pour cette étude le *Lexis Larousse* (LL) et le *Grand Robert* (GR). Nous détaillons ci-dessous la construction de la SL de *amélioration*.

Le GR propose une seule entrée, dont l'article se décompose en trois significations : une première signification générale, que nous conservons pour la construction de la SL ; une seconde signification spécialisée au domaine de l'agriculture (amélioration d'un sol → abonnissement, amendement, bonification) ; une troisième signification réservée aux valeurs mobilières et immobilières (améliorations sur un immeuble → plus-value ; amélioration d'une maison → embellissement, réparation, restauration). Le GR propose deux constructions correctes : *une nette/bonne... amélioration* ; *l'amélioration de qqch*.

Définition : action de rendre meilleur, de changer en mieux, de devenir plus satisfaisant.

→ perfectionnement, progrès (apporter/opérer/réaliser une amélioration ; amélioration morale/spirituelle/intellectuelle)

→ correction, retouche, révision (amélioration d'un travail)

Le LL propose une entrée groupée de – *améliorable*, *améliorateur*, *amélioratif*, *amélioration*, *améliorer* – sans article associé, mais avec une indication de report à l'entrée *meilleur* dont *amélioration* est une sous entrée. Ce sous article présente une acception générale du mot, que nous conservons pour la construction de la SL et une acception spécialisée au domaine technique du bois (amélioration des bois). Dans son acception générale, le sous article ne propose pas de définition, au sens où l'on n'y trouve pas de périphrase synonymique du défini, mais un exemple – *L'amélioration des rapports entre les deux états a amené une détente générale*. Nous nous reportons donc à la définition proposée pour *améliorer*.

Définition : rendre meilleur.

Les deux dictionnaires décrivent *amélioration* comme un nom féminin. Nous reprenons d'ailleurs cette même terminologie dans le tableau de synthèse. Cette unité lexicale appartient à la zone modale de l'axiologique, puisque le mélioratif *meilleur* participe de sa définition (de son noyau), ce qui fait de lui un monovalent positif.

AMÉLIORATION(S)

TRAITS NECESSAIRES DE CATEGORISATION (TNC)		
TYPE GRAMMATICAL	ZONE MODALE	ORIENTATION AXIOLOGIQUE
Nom – Féminin	Axiologique	Monovalent – Positif

ENTITÉS SÉMANTIQUES		ENTITÉS DISCURSIVES
NOYAU	STÉRÉOTYPES	POSSIBLES ARGUMENTATIFS
Processus : rendre meilleur, plus satisfaisant	– Donc perfectionnement, progrès… – Donc correction, retouche, révision…	– Apporter une amélioration – Amélioration morale – Amélioration d'un travail

Tableau 2 : Signification lexicale de amélioration(s)

5. RÉSULTATS

5.1. Liste des collocations lexicales basées sur *amélioration(s)*

Cette sous-section présente le tableau récapitulatif des collocations construites sur *amélioration(s).*

COLLOCATIONS	FREQ.	AUTRES FORMES	FREQ.	TOTAL
Amélioration de la cohérence	1	-Améliorer la cohérence	3	4
Amélioration de la méthode	2	-A. de ces méthodes -A. de méthodes -Améliorer une méthode -cette méthode […] être A.	1 1 1 1	6
Amélioration de la pertinence	1	-Améliorer la pertinence -Améliorer […] la pertinence	2 1	4
Amélioration de la qualité	4	-A. […] de la qualité -Améliorer la qualité -Améliorer […] la qualité -la qualité […] être A.	1 10 1 2	18
Amélioration des performances	13	-A. […] des performances -Améliorations […] performance -Améliorer la performance -Améliorer les performances -Améliorer […] les performances -Améliorer leurs performances -Améliorer notre performance -Améliorer sa performance -les performances […] être A.	4 1 7 38 12 1 2 1 2	81
Amélioration des résultats	12	-A. […] des résultats -A. du résultat -A. en termes de résultats -Améliorations des résultats -Améliorer les résultats -Améliorer […] les résultats -Améliorer leurs résultats -Améliorer nos résultats -ces résultats […] être A.	2 1 1 1 17 9 1 2 2	48
Amélioration des systèmes	3	-A. du système -A. […] de leur système -Améliorer le système -Améliorer les systèmes -Améliorer un système	1 2 2 3 1	12
Amélioration sensible	8	-A. […] sensible -Améliorer de manière sensible -Améliorer sensiblement	2 1 3	14
Amélioration significative	7	-Améliorations significatives -Améliorer […] significative -Améliorer significativement	2 2 6	17
Amélioration substantielle	1	-Améliorations substantielles -Améliorer substantiellement	2 1	4

Tableau 3 : Collocations lexicales contexte droit de amélioration(s)

Nous n'avons relevé aucune collocation basée sur le contexte gauche, à l'exception de *légère amélioration* (Freq. 4sg et Freq. 1pl).

5.2. Analyse de *amélioration(s)*

Nous attribuons le statut de collocation argumentative d'une part, à toutes les collocations lexicales conformes aux constructions grammaticales correctes proposées par l'un ou l'autres des deux dictionnaires[4] ; d'autre part, par un test de substitution sémantique[5], à toutes les collocations lexicales pour lesquelles, en remplaçant la base par les éléments du noyau, l'adjonction au collocatif reste grammaticalement correcte et sémantiquement possible ;

Ex : *amélioration de la cohérence → meilleure cohérence, cohérence plus satisfaisante*

et/ou forme avec le/les stéréotype(s) proposé(s) ou envisageable(s), un bloc d'argumentation interne.

Ex : *meilleure cohérence, cohérence plus satisfaisante → Donc perfectionnement…*

meilleure cohérence, cohérence plus satisfaisante → Donc correction, retouche…

Nous détaillons le raisonnement analytique mené pour la catégorisation des collocations argumentatives, sur la première collocation lexicale recensée dans le tableau de synthèse. Ce même raisonnement est appliqué ensuite sur toutes les collocations lexicales du tableau pour leur répartition en collocation argumentative ou bien inédite.

La collocation *amélioration de la cohérence*, qu'elle soit formée sur le nom ou sur le verbe[6], est conforme aux constructions proposées dans le GR et le LL. Par l'application du test de substitution sémantique, l'adjonction au collocatif :

- reste grammaticalement correcte et sémantiquement possible ;

→ amélioration de la cohérence → meilleure cohérence, cohérence plus satisfaisante

→ améliorer la cohérence → meilleure cohérence, cohérence plus satisfaisante

- forme avec les stéréotypes proposés ou envisageables, un bloc d'argumentation interne.

Nom → meilleure cohérence, cohérence plus satisfaisante →

Donc perfectionnement, progrès… Donc correction, retouche, révision…

Verbe, processus → meilleure cohérence, cohérence plus satisfaisante →

Donc perfectionner… Donc corriger, retoucher, réviser, revoir… Donc évoluer en mieux…

Verbe, résultat → meilleure cohérence, cohérence plus satisfaisante →

Donc perfectionné… Donc corrigé, retouché, révisé, revu…

Par conséquent, nous constatons que la collocation lexicale *amélioration de la cohérence* est la concrétisation d'un PA dans le discours. Il s'agit donc d'une collocation argumentative formant un bloc d'argumentation externe. Elle est intrinsèquement axiologisée, puisque *amélioration* est un monovalent positif, de même que *cohérence*.

Suivant le même raisonnement, nous attribuons le statut de collocation argumentative aux collocations lexicales suivantes : *amélioration de la méthode*, *amélioration de la*

pertinence, *amélioration de la qualité*, *amélioration des performances*, *amélioration des résultats*, *amélioration des systèmes*.

Nous détaillons également le raisonnement analytique mené concernant *amélioration sensible*, transposable à *amélioration significative* et *amélioration substantielle*, car il est à notre sens moins intuitif que le précédent. Ces collocations lexicales correspondent à une construction correcte proposée dans le GR. À la différence des collocations précédentes, par le test de substitution sémantique, l'adjonction au collocatif ne permet pas d'obtenir une structure grammaticale correcte et sémantiquement possible.

Ex : *amélioration sensible* → **rendre meilleur sensible*, **sensible plus satisfaisante*

En revanche, la composition sémantique forme avec les stéréotypes proposés ou envisageables, un bloc d'argumentation interne. Les enchaînements argumentatifs proposés correspondent aux éléments de la SL de *sensible*[7], *significative* et *substantielle*.

Nom → amélioration sensible → rendre meilleur, plus satisfaisante →

Donc plus grand, plus intense, plus important, plus manifeste…

Verbe, processus → améliorer sensiblement → rendre meilleur, devenir plus satisfaisante → Donc changer d'une manière plus appréciable, importante, notable…

Nom → amélioration significative → rendre meilleur, plus satisfaisante →

Donc plus important, plus marquant…

Verbe, processus → améliorer sensiblement → rendre meilleur, devenir plus satisfaisante → Donc changer d'une manière plus importante, plus marquante…

Nom → amélioration substantielle → rendre meilleur, plus satisfaisante → Donc plus fondamental, plus essentiel…

Verbe, processus → améliorer substantiellement → rendre meilleur, devenir plus satisfaisante → Donc changer d'une manière plus fondamentale, plus essentielle…

Par conséquent, ces trois collocations correspondent à la concrétisation d'un PA dans le discours. Il s'agit donc de collocations argumentatives formant des blocs d'argumentation externe, intrinsèquement axiologisées, puisque *amélioration* est un monovalent positif, de même que *sensible*, *significative* et *substantielle*, qui fonctionnent tous comme des intensifieurs.

6. CONCLUSION

Les analyses menées jusqu'à présent sur les nœuds *amélioration(s)*, *connaissance(s)*, *savoir*… prouvent l'existence de collocations argumentatives et l'absence de collocation inédite. Toutes les collocations argumentatives, sont intrinsèquement axiologisées, ici positivement, conformément à l'orientation axiologique de *amélioration*. Ces données confortent en partie nos hypothèses quant à la concrétisation des collocations lexicales sous la forme de PA. Toutefois, l'existence de collocations inédites reste hypothétique, avec elle, la faculté de la base à conserver son potentiel axiologique et le principe d'axiologisation extrinsèque par relais contextuel. De toute évidence, cette étude remet en cause le caractère considéré comme arbitraire des collocations et ouvre une perspective d'explication possible de leur principe de composition en terme de bloc d'argumentation, par relation sémantico-discursive motivée.

7. BIBLIOGRAPHIE

Benson M., **Benson E.** & **Ilson R.** 1986. *The BBI Dictionary of English Word Combinations*. Amsterdam, Benjamins (2ème ed, 1997).

Carel M. & **Ducrot O.** 1999. Le Problème du paradoxe dans une sémantique argumentative, *Langue française*, sept. 1999, p. 6-26.

Daille B. & **Williams G.** dir. 2001. *Collocation: computational extraction, analysis and exploitation*. Workshop at the 39th Annual Meeting and 10th Conference of the European Chapter of ACL.

Dubreil E. 2005 (à paraître). 'Principes de constitution du corpus TAL(N) dédié à une analyse sémantico-discursive', dans *Recueil des données en Sciences du langage et constitution de corpus : données, méthodologie, outillage. 2ème colloque Jeunes Chercheurs COL'DOC 2005*, Paris X - Nanterre, 16-17 juin 2005 *(à paraître).*

Dubreil E. 2005 (à paraître). 'Etude des collocations épistémiques en corpus spécialisé : de la théorie à l'empirie…' dans Vergely, P. (coord.) *Rôle et place des corpus en linguistique*. Actes *des Journées d'études Doctorants et jeunes chercheurs (JETOU)*, 1-2 juillet 2005, Université de Toulouse Le Mirail. Toulouse : Ch.F.-P.

Firth J.R. 1957. *Papers in Linguistics 1934-1951*. Oxford, Oxford University Press.

Galatanu O. 2000. 'Langue, discours et systèmes de valeurs', dans Suomela-Salmi (eds) 2000. *Curiosités linguistiques*, Université de Turku.

Galatanu O. 2004. Sémantique des "possibles argumentatifs" et axiologisation discursive, dans Bouchard (I.), *Représentation du sens linguistique II*, Paris Louvain la Neuve Duculos.

Galatanu O. 2005. Sémantique et élaboration discursive des identités. « L'Europe de la connaissance » dans le discours académique, conférence plénière, Actes de la conférence *international-cultural and cross-lingusitic Perspectives on Academic Discourse*, May 2005, University of Turku, Finland, school of Classical and Romance Languages *(à paraître).*

Grossmann F. & **Tutin A.** dir. 2003. Les collocations : analyse et traitement, coll. *Travaux de recherches en linguistique appliquée*, Paris : Éditions De Werelt.

Haussmann F.J. 1997. Tout est idiomatique dans les langues, In M. Martins-Baltar (éd.). *La locution entre langues et usages*. Paris, ENS éditions, 277-290.

Kleiber G. 1999. *Problèmes de sémantique, La polysémie en question*, Nancy, Presses Universitaires du Septentrion.

Mel'čuk I. 1984, 1988, 1992, 1999. *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques*. Vol. I, II, III, IV. Montréal, Presses de l'Université de Montréal.

Van der Wouden T. 1997. *Negative contexts*. Routledge, London.

Putnam H.1994. *Le réalisme à visage humain*, Paris, Armand Colin.

Williams G.1998. Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles, *International Journal of Corpus Linguistics*, 3(1), 151-171.

Williams G. 2001. Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique, Lille : Presses Universitaires de Septentrion.

Williams G. 2003. Les collocations et l'école contextualiste britannique, dans Grossmann (F.) et Tutin (A.) eds. Les collocations : analyse et traitement, coll. Travaux de recherches en linguistique appliquée, Paris : Éditions De Werelt.

8. NOTES

1. Logiciel développé par Mike Scott, téléchargeable gratuitement en version limitée à l'adresse suivante : http://www.lexically.net/wordsmith/

2. Logiciel développé par Jean Véronis, téléchargeable gratuitement en version limitée à l'adresse suivante : http://www.up.univ-mrs.fr/~veronis/logiciels/Contextes/index-fr.html

3. EURODICAUTOM : http://europa.eu.int/eurodicautom/Controller

Grand Dictionnaire Terminologique :
http://w3.granddictionnaire.com/btml/fra/r_motclef/index1024_1.asp
4. À noter que ce critère est redondant, puisqu'il intervient déjà dans la sélection des candidats collocation, en qualité de paramètre définitoire de ces mêmes entités.
5. Le test de substitution sémantique nécessite parfois le recours à la définition du collocatif pour s'assurer de la convenance grammaticale et de la cohérence sémantique ; nous utilisons au besoin le TFLi, http://atilf.atilf.fr/tlf.htm
6. La plupart des collocations présentent des variantes construites sur le verbe (*amélioration de la cohérence*, *améliorer la cohérence*), nous avons donc également construit la SL du verbe *améliorer*. Elle est identique à celle du nom, à la seule différence qu'il existe deux noyaux, selon qu'il s'agisse du processus ou du résultat. Aussi, nous ne présentons pas la synthèse de la SL de *améliorer*.
7. Lorsque la collocation lexicale présente plusieurs formes (amélioration sensible, améliorer sensiblement), nous avons tenu compte des variations éventuelles des éléments de signification lexicale du collocatif (entre sensible et sensiblement par exemple).

CONSTRUCTIONS DE CLASSES SEMANTIQUES NON PRESUPPOSEES

Nicolas Gagean
François Rousselot
LGECO - INSA

1. INTRODUCTION

La construction de classes sémantiques est un thème récurrent dans le domaine du TAL. L'approche distributionnelle permet de créer des classes sémantiques de mots sur la base des contextes linguistiques de ces mots. En d'autres termes, le fait d'avoir des contextes communs est une indication de proximité sémantique pour deux mots. Le degré de similarité est fonction des contextes partagés. Les différentes approches, essentiellement celles qui sont supervisées, font souvent le postulat de l'existence au préalable des classes. Il s'agit plus dans ce cas de ranger les éléments trouvés dans des classes pré-établies ou au moins présupposées, plutôt que de faire réellement émerger des ensembles de mots sémantiquement proches. À l'inverse, parmi les méthodes peu ou pas supervisées, nous pouvons noter l'existence d'outils utilisés pour la construction de terminologies (Nazarenko *et al*, 1997), d'ontologies (Bourigault, 2002), de thésaurus (Grefenstette, 1994) ou encore pour l'extraction de connaissances (Faure & Nedellec, 1999).

Nous nous proposons de faire une étude prospective sur corpus avec pour objectif l'extraction de classes de termes dont la proximité tient plus de l'usage que du sens. C'est l'examen des distributions des termes qui nous servira de critère de regroupement. Nous procédons en plusieurs temps : d'abord avec des critères très restrictifs afin de limiter le nombre de classes et de réduire le bruit. Ensuite, nous essayons de traiter le silence : il faut adapter les critères de regroupement afin de rassembler plus d'éléments (introduction de seuil, etc.). Il est évident qu'à un certain moment, nous sommes obligés de faire appel à un jugement extérieur car l'assouplissement des critères a pour corollaire l'introduction d'éléments non pertinents dans les classes.

Plusieurs manipulations permettent d'obtenir des classes initiales (430 classes dont 130 sont jugées pertinentes), et divers aménagements montrent l'influence du contexte sur les classes obtenues et leur pertinence. Les critères de regroupement (nombre de contextes pris en compte, fréquence, introduction d'un seuil de comparaison) sont réajustés afin de faire évoluer le bruit, le silence ainsi que la taille des classes. Nous pouvons réduire le nombre de classes en augmentant le taux de classes intéressantes, ou au contraire augmenter le nombre total des regroupements pour faire émerger parmi elles des classes plus conséquentes.

2. APPROCHE THÉORIQUE

Le point majeur de notre travail est que nous ne présupposons pas *a priori* l'existence des classes qui seront créées. Ce rejet du concept est une approche objectiviste, qui s'oppose au mentalisme ou cognitivisme généralement en vogue, et s'inspire de théories sémiotiques.

Nous nous appuyons sur les travaux de Frath (1997) qui, à la suite de Wittgenstein et de Pierce, conteste l'attribution de contenus intrinsèques aux signes (aux mots). Il réfute l'existence réelle des catégories car elle suppose des contenus en dehors de la langue et préconise de s'appuyer sur l'usage pour définir le sens. En résumé, la sémantique n'est pas innée mais s'acquiert par le biais de l'usage. Cet usage s'exprime par des dénominations pré-construites et des signes complexes construits (appelés *interprétants*). Les dénominations nomment les objets et pré-existent à notre apprentissage. Les signes complexes (représentés ici par les contextes) naissent de l'usage et participent au lien référentiel entre le signe et l'objet, qui est assimilable à une opération de catégorisation. Par conséquent, il est fondé de penser que les catégories s'acquièrent aussi grâce à l'usage.

La thèse de la catégorisation d'un point de vue sémiotique telle qu'elle est défendue par Frath s'appuie sur le rejet de Wittgenstein de la conception sociologique de la vérité. Frath affirme que les mots réfèrent par nature et se demande comment nous pouvons les utiliser pour désigner d'autres référents. Cela nous ramène au problème de la catégorisation, avec les solutions empiriques et réalistes, solutions que rejette Wittgenstein.

Le modèle empirique est celui des Conditions Nécessaires et Suffisantes (CNS), selon lequel les choses se composeraient de qualités que l'on peut nommer. L'appartenance à une catégorie se fait en termes de CNS. C'est une approche qui peut sembler naturelle mais dont la simplicité ne résiste pas à la complexité des phénomènes sémantiques comme la polysémie et la synonymie. Le positivisme logique a consisté à tenter de remédier à ces « défauts » des langues naturelles par la création d'une langue scientifique fondée sur la logique. Cela a abouti aux travaux des terminologues.

Le modèle des CNS est compositionnel : les choses se composent d'autres choses moins complexes et ainsi de suite jusqu'à des éléments indécomposables. Il pose par conséquent le problème des primitives sémantiques : il existerait des indécomposables sémantiques universelles, obtenues en comparant des échantillons de langues. Si un mot est difficile à définir, on fait l'hypothèse qu'il s'agit d'une primitive et on regarde s'il revient pour chacune des langues de son échantillon. Cette approche comporte de nombreuses difficultés, comme l'arbitraire des primitives, le fait qu'il faille disposer d'un locuteur de chaque langue (ce qui est pratiquement impossible). Elle se révèle aussi contre-intuitive sous certains aspects : les déictiques sont absents ainsi que les mots logiques comme *et* et *ou*. L'hypothèse des primitives sémantiques cherche à donner une base empirique à toutes les approches componentielles du sens, mais elle n'explique rien quand à la nature des contenus.

La théorie du prototype représente le modèle réaliste, dans lequel on se réfère à une représentation mentale. Ce n'est plus le monde réel qui compte mais la manière dont nous nous le représentons. Chaque locuteur se représente les catégories au moyen d'occurrences particulièrement représentatives. Le mode d'appartenance change et se fait en fonction de la ressemblance avec ces exemplaires typiques. Ceux-ci sont réels dans la version standard de la théorie. Ils perdent leur contrepartie réelle dans la version étendue du prototype. Dans les deux cas nous avons à faire à une représentation mentale du sens d'un mot, ce qui oblige les mots à exprimer un contenu.

Pour Wittgenstein, la vérité, comme le langage, est une entité qui cesserait d'exister avec la fin de l'homme. Et c'est par refus de tout essentialisme que Wittgenstein rejette ces deux solutions. Celles-ci supposent que les catégories sont réelles, qu'elles existent indépendamment de l'homme. C'est là l'approche cognitiviste, qui postule l'existence de concepts. Nous aurions dans nos têtes des multitudes de sens que nous exprimons à l'aide

de signes et qui réfèrent à la réalité. Mais bien qu'elles soient riches d'enseignements sur bien des points, aucune théorie cognitiviste n'a pu expliquer ce qu'est un concept. Elles se sont attachées à tenter de décrire le fonctionnement de quelque chose d'insaisissable, et somme toute de très hypothétique.

Nous rejoignons Frath, lorsqu'il conteste, à la suite de Wittgenstein, l'attribution de contenus intrinsèques aux signes, et donc les théories qui expliquent le lien entre les contenus et la vie. Le corollaire est le rejet de l'existence réelle des catégories car elle suppose des contenus en dehors de la langue. Nous préconisons de s'appuyer sur l'usage pour définir le sens, et nous présentons ici la formulation de Frath d'une conception du langage et des catégories qui correspond à cela.

Selon Pierce (1978), on ne peut établir un rapport cognitif avec un objet extérieur que par le biais d'un signe. En d'autres termes, toute connaissance est liée à un signe. Un signe peut déterminer un autre signe à renvoyer à un objet auquel lui-même renvoie. La sémiotique de Pierce introduit l'objet comme entité extérieure que nous avons la capacité de rendre significative pour nous. La fonction du signe est de désigner l'objet, il n'apporte pas d'informations dessus. Frath distingue le *representamen* ou dénomination, de l'interprétant que constituent le syntagme et l'énoncé. La dénomination présuppose l'existence de l'objet dénommé. Kleiber quant à lui oppose la dénomination, préconstruite (*pomme de terre*) à la désignation, construite (*légume avec lequel on fait des frites*). Frath et Kleiber s'accordent à voir en la catégorie la référence d'une dénomination. Dans cette optique, le sens lexical se définit en termes d'appartenance à une catégorie. Par contre leurs explications quant à ce fait divergent. Kleiber pense que c'est l'unicité de la dénomination qui permet l'appartenance à une catégorie, point de vue que rejette Frath en contestant cette unicité de la dénomination, dans la mesure où elle n'est pas systématique (*carte bleue*). Selon Frath, « *les catégories sont fondées sur le signe qui rassemble des occurrences référentielles sous une seule dénomination* ». La catégorie est « *un ensemble d'objets provisoirement considérés comme ressemblants et qui acquiert une existence sociale, dont la dénomination est le signe matériel* ».

Frath soutient que l'objet est porté à la connaissance en même temps que sa dénomination. Si on admet cela, force est de reconnaître que la capacité à dénommer et à catégoriser naît de l'usage. Dans ce sens, le signe ou *representamen* ou encore dénomination, l'objet et le lien qui les unit sont donnés et préconstruits. Ce sont les connaissances à propos de l'objet qui sont à construire, et qui se composent d'autres *representamen*. Ce sont des signes interprétants. C'est la production d'interprétants qui entretient un jeu créatif entre le signe opaque et l'objet. Frath définit la dénomination comme public, stable, pré-construite et opaque tandis que la désignation est éphémère, construite et transparente.

Cette théorie, qui dénie aux signes la fonction de porteur de sens, peine pourtant à expliquer le lien entre dénomination et objets abstraits, comme *liberté* par exemple. Il n'existe pas un objet *liberté* auquel ce signe fait référence, mais il y en a autant qu'il y a de locuteurs. Ici, au contraire des objets concrets, les interprétants ne permettent pas de spécifier un objet commun, public. Dans ce cas, comment concevoir un lien public pré-construit pour un objet qui se construit de façon unique selon l'expérience et l'usage ? On en arrive à la contradiction suivante, où la dénomination, en dehors d'interprétants, n'est plus le lien référentiel entre un objet et un signe, mais le lien entre un signe et une catégorie mentale. Le signe redevient porteur de contenu.

En résumé, comme Frath nous refusons de faire reposer l'explication linguistique sur l'existence de règles, de primitives ou de concepts indépendants. Nous réfutons l'existence de contenus stables qui pré-existent à l'énoncé. La sémantique n'est pas innée comme peut l'être dans une certaine mesure la grammaire, mais s'acquiert par le biais de l'usage. Cet usage s'exprime par des dénominations pré-construites et des signes complexes construits. Les dénominations nomment les objets et pré-existent à notre apprentissage. Le lien entre la dénomination et l'objet est assimilable à une opération de catégorisation. Par conséquent, il est fondé de penser que les catégories s'acquièrent aussi grâce à l'usage.

Mais refuser aux catégories un statut indépendant de la langue ne revient pas à les nier complètement pour autant. Elles sont indispensables à l'acte langagier et leur utilité n'est plus à démontrer. Simplement les catégories sont artificiellement générées par des besoins cognitifs et n'ont pas de fonction en dehors de cela. Nous rejoignons pleinement Frath lorsqu'il affirme qu'« *il n'y a pas de catégories dans la nature, il n'y a que des occurrences que nous rassemblons sous des dénominations* ». C'est donc la dénomination qui crée la catégorie, et non l'inverse.

Nous venons de voir que c'est l'apprentissage qui permet de lier l'objet et la dénomination. Par ailleurs, la dénomination, par le biais de la catégorie qu'elle crée de par son existence, rassemble une pluralité d'objets assez analogues. Mais cet apprentissage ne permet pas d'expliquer tous les liens qui existent entre une dénomination et tous les objets qu'elle désigne. Par exemple, nous pouvons reconnaître dans un objet jamais vu auparavant qu'il s'agit d'une table, sans que personne ne porte ce lien à notre connaissance. C'est une opération d'analogie qui nous permet de ranger cet objet dans la catégorie *table* et de le nommer ainsi. Comment expliquer cela, dans une théorie qui rejette l'existence d'un concept *table* ? En d'autres termes, avec quoi l'objet peut-il établir une ressemblance ?

Le phénomène de polysémie met en évidence le même problème. Considérant son exemple de la souris (le rongeur) et de la souris d'ordinateur, Frath constate que « *l'établissement d'une ressemblance, même partielle, avec un membre d'une catégorie suffit à permettre un nouvel usage* » et ajoute que « *ce phénomène de polysémie lexicale est la création de la créativité linguistique* ». Certes, mais il n'explique pas clairement sur quelle base se fait cette ressemblance. Dans ce cas comme dans le précédent, il fait reposer la capacité de la dénomination à regrouper sur la seule présupposition d'existence de la catégorie. Il montre avec justesse que la dénomination génère une analogie entre les objets qu'elle regroupe, mais cela n'explique pas comment un nouvel usage pour la dénomination peut s'établir, ou encore ce qui autorise un nouveau lien entre la dénomination et un objet vu pour la première fois. Sans rejeter aucune de ces explications, nous allons éclaircir cet aspect en allant plus avant dans cette approche sémiotique, avec toujours comme point de mire l'usage.

Personne ne nous contredira lorsque nous affirmons que l'analogie, que ce soit pour la polysémie ou pour le regroupement d'éléments dans une même catégorie, repose sur une ressemblance physique et/ou fonctionnelle, le problème étant l'établissement de cette ressemblance. Il est donc question de l'apparence et de la fonction des objets. Or nous avons vu précédemment que les connaissances de ce type à propos des objets n'ont pas d'existence en elles-mêmes. Elles s'expriment et n'existent qu'au moyen des signes interprétants. Frath explique que « *le contenu sémantique des unités lexicales est constitué de signes interprétants produits à propos de l'objet* ». Si nous voyons une voiture bleue neuve de grosse cylindrée, c'est le lien acquis lors de l'apprentissage qui nous permet de ranger cet objet dans la catégorie *voiture*, et les connaissances quant à ses caractéristiques

physiques que sont sa couleur, son âge et sa puissance existent en nous grâce aux signes interprétants *bleu*, *neuve* et *grosse cylindrée*, ces signes eux-mêmes étant issus de l'apprentissage. Nous postulons alors que la ressemblance dont nous parlions précédemment s'établit sur les signes interprétants associés aux objets qui sont regroupés dans une même catégorie. De même le phénomène de la polysémie s'explique selon ce point de vue par le partage d'un certain nombre de signes interprétants.

Certains seront tentés de reprocher à cette explication qu'elle ne fait que déplacer le problème, en passant de caractéristiques sémantiques aux signes. À la place d'une analogie forte ou partielle de parties de sens, on a une analogie de signes interprétants. À cela nous répondrons que l'existence des signes que sont les mots n'est pas sujette à caution, comme peut l'être celle des sèmes ou autres concepts sémantiques indépendants. Ce point de vue, s'il n'explique pas tout, a l'avantage de reposer sur des bases solides et concrètes que sont les signes, puisqu'il fait abstraction des règles, concepts ou autres présupposés cognitivistes.

Nous avons établi avec Frath que la catégorie est la référence de la dénomination et qu'elle constitue un regroupement d'une pluralité d'objets plus ou moins ressemblants sur la base de l'usage, c'est-à-dire des signes interprétants qui sont produits à propos de ces objets. Mais qu'entendons-nous par usage ? Qu'est-ce qui permet à une dénomination d'être utilisée dans tel emploi plutôt que dans tel autre ? La question est d'importance, car le sens lexical de cette dénomination en dépend.

Pour expliquer certaines contraintes syntagmatiques, Frath ne retient pas le mécanisme de la coercition de type de Pustejovsky en assurant que le problème ne vient pas d'un mécanisme de sélection de traits sémantiques. De plus, ce mécanisme n'est somme toute qu'une règle, par conséquent réfutée par principe dans cette approche sémiotique. Frath explique que « *les énoncés sont des signes interprétants à propos d'objets* » qui sont contraints non pas par des contenus sémantiques pré-existants mais par l'usage. Il réduit les règles à « *un discours sur la langue* » dans la mesure où « *elles ne sont pas la cause de la langue* ». Ainsi, il postule l'existence d'une grammaire des signes qui n'est pas dirigée par des règles mais bien par l'usage. La grammaire d'un signe représente les possibilités de liens entre ce signe et des objets dénommés dans le contexte proche.

Pour conclure sur cette approche sémiotique et pour résumer ce qui a été dit, nous reprenons la définition que propose Frath des trois parties qui composent le sens lexical d'une dénomination. En premier lieu la dénomination pointe vers une catégorie, qui existe de par l'usage de cette dénomination. C'est un aspect essentiel car cette catégorie regroupe une pluralité d'objets (référentiels ou linguistiques), ce qui correspond au propos de notre étude. La deuxième partie du sens lexical est constituée par les connaissances exprimées par les signes interprétants sur ces objets. Nous avons vu que ces connaissances participent au regroupement des objets. Le dernier élément sémantique de la dénomination est sa grammaire, c'est-à-dire sa capacité à entrer en relation syntagmatique avec d'autres dénominations. C'est sur la base de ce dernier point que nous mènerons notre travail, qui consiste à étudier les possibilités de regroupements de mots en s'appuyant sur l'étude de leurs contextes dans des énoncés.

Le TAL peut difficilement s'appuyer sur l'approche conceptuelle car la notion même de concept varie d'une théorie à l'autre. Il est impossible d'en donner une définition valide et précise dans la mesure où elle reposerait entièrement sur la subjectivité de son auteur. Tout ce qui a été développé dans ce sens a considérablement fait évoluer la linguistique et les recherches sur la cognition, mais se révèle peu exploitable en TAL. Les théories

conceptuelles avancent des postulats qui manquent de critères formels, et sont par là même difficiles à être reproduites artificiellement. Introduire manuellement des concepts ou des définitions dans un système de TAL revient à l'appauvrir en forçant à un parti pris. Le TAL doit s'affranchir de tout point de vue subjectif pour parvenir à des résultats indépendants et réutilisables. En résumé, les concepts n'existent pas en eux-mêmes mais sont toujours les concepts de quelqu'un, tandis que les mots en tant que signes sont les mêmes pour tout le monde, et si l'usage qui en est fait peut varier selon les individus, il obéit néanmoins à des contraintes fortes qui sont autant de critères exploitables.

3. TRAITEMENT

Le corpus Ménélas a été traité manuellement de la façon suivante : tous les substantifs ont été extraits. Pour chaque substantif, nous avons relevé sa fréquence d'apparition dans le corpus et ainsi que tous les éléments lexicaux du type nom, verbe ou adjectif qui sont en relation de dépendance syntaxique avec ce substantif. La fréquence de chacune des relations de ce type a aussi été relevée. Une lemmatisation sommaire a été appliquée : le nombre et le genre ont été ramenés au masculin singulier, et les verbes sont tous à l'infinitif, ceci afin d'optimiser les possibilités de comparaisons des contextes. Par exemple, le substantif accélération a une seule occurrence, avec un contexte gauche et un contexte droit.

> Au cours de la crise angineuse, on note fréquemment une élévation de la tension artérielle et une ***accélération*** de la fréquence cardiaque.

On traduit cela par la notation suivante :

	noter	
accélération	1	
	de fréquence	

En accord avec l'approche théorique adoptée et décrite précédemment, nous avons choisi de comparer les éléments deux à deux, plutôt que de tous les comparer à une seule entité englobante. En effet nous ne désirons pas baser notre étude sur la présupposition de l'existence de classes pré-établies. Supposons que le critère de regroupement soit que les contextes des entités sont inclus dans ceux de l'entité E pour appartenir à la classe C. Cela signifierait que l'on présuppose qu'il existe une classe C dont le prototype, ou meilleur représentant, serait l'entité E. C'est une approche cognitiviste à laquelle nous n'adhérons pas.

3.1. Première analyse

Dans un premier temps, la procédure de regroupement est très stricte afin de n'obtenir que des classes pertinentes, même si elles sont peu nombreuses. Nous considérons que tous les substantifs extraits du Ménélas sont les sommets d'un graphe. Un arc entre deux sommets indique qu'il y a une relation d'inclusion entre les deux sommets. Une relation d'inclusion entre deux éléments A et B (ou sommets ou entités) signifie soit que tous les contextes de A sont aussi des contextes de B, soit que tous les contextes de B sont aussi des contextes de A. Le graphe est non-orienté. Les classes sont les cliques maximales du graphe, c'est-à-dire les sommets qui sont tous reliés les uns aux autres. Cela signifie que, pour une classe donnée, chaque élément de la classe entretient une relation d'inclusion avec tous les autres éléments de la classe.

Avec cette première passe, nous obtenons 550 classes dont 122 sont jugées pertinentes. Elles comprennent jusqu'à 6 éléments, mais il y a une majorité de couples. Si certaines classes sont intéressantes, il reste toutefois beaucoup de bruit. De plus, certaines classes gagneraient à être regroupées.

Exemples de classes intéressantes :

- *tabac / alcool*
- *distorsion / agrandissement*
- *scintigraphie / biologie*
- *scintigraphie / radiographie*
- *jour / surlendemain*
- *conservation / altération*
- *métaraminol / lidocaïne / atropine*
- *décision / avis*

Exemples de cas non pertinents :

- *fatigue / boulanger*
- *prothrombine / positif / morbidité*

Il y a 30 regroupements dans lesquels on retrouve les éléments *fatigue* et *boulanger*. Le seul contexte avec lequel ils apparaissent est le prédicat *être*, qui est extrêmement fréquent et qui est très pauvre en information contextuelle. Dans le deuxième cas, c'est le contexte *taux de* qui a permis ce regroupement et qui n'est pas suffisamment spécifique.

En conclusion pour cette première analyse, le fait de ne pas prendre en compte les expressions verbales figées amène à des contextes simplifiés comme *avoir* ou *être*. En l'état, ces contextes ne donnent pas d'indication sémantique permettant des regroupements pertinents et ne doivent par conséquent pas être pris en compte.

3.2. Deuxième analyse

La seconde passe est un affinement de la première : le procédé est le même, à la différence que l'on ne retient plus avoir et être comme contextes. Cela permet d'obtenir 423 classes dont 130 sont jugées pertinentes. Etant donné la rigueur des critères, les classes se composent quasiment toutes de 2 termes seulement. Ensuite nous avons tenu compte pour chaque dénomination du nombre de ses contextes ainsi que de sa fréquence. Le fait de restreindre les critères de regroupement (notamment au niveau des contextes) fait diminuer considérablement le nombre de classes, tout en augmentant la proportion de classes pertinentes (66 classes dont 53 pertinentes, puis 11 classes dont 10 pertinentes).

Avant de passer à la troisième analyse avec introduction de seuils, nous avons examiné le volume des contextes et la fréquence des termes qui constituent les classes pour expliquer pourquoi les deux tiers des classes restent indésirables. L'analyse des contextes est importante pour distinguer les contextes qui permettent de regrouper des classes intéressantes des contextes à l'origine de regroupements inintéressants. Après examen, il apparaît que le caractère répétitif du corpus fait que peu de contextes différents interviennent dans un cas comme dans l'autre. De plus l'intérêt des contextes ne semble pas lié à un type syntaxique ou sémantique particulier.

À ce stade, nous supposons que les termes de faible occurrence sont la cause des regroupements non désirés. Ces éléments sont des signes flous car ils possèdent peu

d'interprétants, ce qui leur permet d'apparaître dans diverses classes. La conclusion est qu'il faut disposer de suffisamment de connaissances sur un terme pour pouvoir le classer de façon satisfaisante. Un signe est classé sur la base de ses interprétants. Avec notre procédure, plus il a d'interprétants, moins il peut apparaître dans des classes, ce qui renforce leur homogénéité. Ceci nous amène à nous intéresser à la fréquence des termes, ainsi qu'aux nombres de contextes différents (c'est-à-dire d'interprétants) de chaque terme.

Si nous ne retenons que les éléments de fréquence supérieure à 1 et qui ont plusieurs contextes différents, nous obtenons 10 classes :

- *débit / cholestérol*
- *biologie / aortographie*
- *biologie / bilan*
- *Sectral / bêtabloquant*
- *conduction / cinétique*
- *contrôle / biologie*
- *coronarographie / biologie*
- *Diafusor / Corvasal*
- *Sectral / IEC*
- *scintigraphie / biologie*

La seule classe non pertinente est *débit / cholestérol*. Le pourcentage de classes intéressantes est très satisfaisant. Ainsi nous constatons que chaque terme doit être suffisamment « décrit » pour pouvoir être classé judicieusement. Comme les connaissances dont nous disposons sur ce terme sont les contextes avec lesquels il apparaît, seuls les termes possédant plusieurs contextes doivent être considérés.

3.3. Troisième analyse

Enfin, nous avons voulu agrandir la taille des classes. Pour cela, nous avons introduit un seuil modulable dans la comparaison des contextes : pour un seuil de 50 %, on divise le nombre de contextes par deux. Par exemple, si un terme possède 6 contextes, le seuil sera 3. Il faudra donc que la dénomination ait au minimum 3 contextes en commun avec le substantif auquel il est comparé pour pouvoir apparaître dans la même classe. En jouant avec ce seuil, nous produisons certes du bruit, mais nous obtenons aussi des classes très pertinentes et composées de plusieurs termes comme l'illustrent les deux classes suivantes (le chiffre indique la fréquence du terme dans le corpus) :

- *sténose 183 / réocclusion 4 / resténose 22 / plaque 15 / occlusion 32 / lésion 104 / irrégularité 9*
- *ventriculographie 31 / scintigraphie 17 / examen 67 / épreuve 57 / coronarographie 237 / angiographie 9*

Le nombre de classes obtenues est fonction du seuil : avec un seuil de 90 %, nous avons 17 classes tandis qu'avec un seuil de 10 %, nous obtenons 336 classes. Ces regroupements ne sont certes pas tous à retenir, cependant cela permet de faire émerger des classes très intéressantes comme :

- *présence 20 / persistance 7 / apparition 9 / absence 24*
- *élévation 12 / aggravation 2*

- *tabagisme 11 / hypertension 7 / diabète 5*
- *semaine 15 / mois 34*
- *technique 24 / méthode 6 / examen 67*
- *revascularisation 28 / pontage 51 / angioplastie 86*

3.4. Vers des classes … d'usages

Examinons une des classes présentées ci-dessus : *tabagisme / hypertension / diabète*. Ces trois termes sont fort différents et ce regroupement n'aurait pas sa place dans une taxinomie conceptuelle classique. Or dans notre étude c'est l'usage qui est fait de ces termes qui permet de les regrouper.

Par l'étude des interprétants, on peut isoler des classes sémantiques véritablement créées par l'usage, comme celle des facteurs de risque d'un point de vue médical. Il apparaît impropre d'utiliser le terme de catégorie, ou de catégorisation, du moins dans son sens aristotélicien. Nous ne mettons pas à jour quelque chose qui existe en soi, de façon universelle et intemporelle, et inhérente à la langue. Ces éléments ne sont réunis que par l'usage qui en est fait dans ce corpus. Il n'y a aucune règle formelle qui permette de les regrouper, si ce n'est leur aptitude à *fonctionner* comme facteurs de risque. De telles classes connaissent un certain flou, rassemblent des éléments par ailleurs très hétérogènes, et ne sont pas valables de façon universelle et intemporelle : elles sont fonction du corpus, du domaine, du moment, etc.

3.5. Un cas particulier d'analyse

La distribution de certains termes les fait apparaître dans des classes, et ce à première vue de façon impropre, comme dans les 4 classes suivantes : *vomissement / constitution*, *vomissement / élévation*, *vomissement / récidive*, *vomissement / survenue*. En fait ce n'est pas si inexact que cela, si on examine les substantifs qui sont complément du nom des deuxièmes termes de ces classes :

- *constitution de sténose, d'infarctus*
- *élévation de tension*
- *récidive d'infarctus, d'angor, d'angine*
- *survenue de bradycardie, d'infarctus, de douleur*

Les termes comme *constitution*, *élévation*, *récidive* et *survenue* ont peu de sens en eux-mêmes, et surtout ils sont référentiellement dépendants du nom qui les complémente. On les appelle des faux-noms. Dans ce cas de figure, N2 est la tête sémantique du schéma N1 de N2.

4. EN GUISE DE CONCLUSION

Le problème est que, au-delà des critères distributionnels utilisés pour regrouper des termes, il n'existe pas de critères formels permettant de valider les regroupements ainsi obtenus. Comment évaluer la pertinence de ces classes ? Cela se fait selon l'intuition de la personne qui effectue cette tâche. Tandis que la pertinence de certaines classes est évidente (*ventriculographie* / *scintigraphie*), d'autres sont beaucoup plus conjoncturelles (*cf.* le cas des *antécédents médicaux*). Une même personne arrivera à deux résultats différents si elle doit se prononcer sur la validité d'un grand nombre de classes à deux moments différents.

On serait tenté de valider une classe dès lors qu'il y a un terme qui permet de la recouvrir. Par exemple les *examens médicaux* pour *coronarographie / ventriculographie / ECG / scintigraphie* / etc. Mais d'autres cas ne vont pas dans le sens de cette approche ontologique : dans un domaine particulier, des termes gagnent à être rapprochés même s'ils ne peuvent pas entrer dans la même catégorie, comme *image / brillance*.

En conclusion, les résultats de cette étude confirment le postulat de départ, à savoir que ce sont des similitudes de comportement entre des éléments qui permettent de les réunir dans une même classe, et non l'appariement avec un concept existant ou non. Si on rejette les théories mentalistes comme le prototype et les Conditions Nécessaires et Suffisantes (CNS), ainsi que la présupposition de l'existence de classes *a priori*, on opte pour l'idée de classes floues, fluctuantes, instables, à l'image de l'usage qui permet de les construire.

5. REFERENCES

Bourigault D. 2002. 'Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus' dans *Actes de la 9e conférence annuelle sur le Traitement Automatique des Langues (TALN 2002).* Nancy.

Faure D. & Nedellec C. 1999. 'Knowledge acquisition of predicate argument structures from technical texts using Machine Learning : the system ASIUM' dans *Proceedings of the 11th European Knowledge Acquisition Workshop (EKAW 1999).* Dagstuhl (Allemagne).

Frath P. 1997. *Sémantique, référence et acquisition automatique de connaissances à partir de textes.* Thèse de Sciences du Langage. Université Marc Bloch. Strasbourg.

Gagean N. 2003. 'Corpus et classes d'objets' dans *Scolia n°16*. Strasbourg : Presses de l'Université de Strasbourg.

Grefenstette G. 1994. *Explorations in automatic thesaurus discovery*. Boston : Kluwer Academic Publishers.

Kleiber G. 1990. *La sémantique du prototype*. Paris : PUF.

Nazarenko A., Zweigenbaum P., Bouaud J. & Habert B. 1997. 'Corpus based identification and refinement of semantic classes' dans *Journal of the American Medical Informatics Association n°4.*

Pierce C. 1978. *Ecrits sur le signe*. Paris : Seuil.

CARTOGRAPHIE DE CORPUS POUR L'ÉTUDE DE MÉTAPHORES CONCEPTUELLES

Thibault Roy, Stéphane Ferrari, Pierre Beust
GREYC – CNRS UMR 6072 – Université de Caen – 14032 CAEN Cedex, France

1. INTRODUCTION

Dans cet article, nous présentons des expériences récentes menées dans le cadre du projet IsoMeta portant sur l'étude de métaphores conceptuelles au sens de (Lakoff et Johnson, 1980). Ces expériences consistent à appréhender visuellement la répartition de lexiques de domaines sources de métaphores conceptuelles dans un corpus d'articles journalistiques traitant d'économie. Dans la partie 2 de cet article, nous abordons les enjeux du projet ainsi que l'intérêt d'une étude de certaines métaphores conceptuelles à l'échelle de la collection de textes. La partie 3 décrit le corpus de travail et les ressources lexicales représentant les domaines sources des métaphores conceptuelles étudiées. La partie 4 de cet article présente différentes vues sur le corpus prenant en considération les domaines étudiés. Ces vues ont été obtenues à l'aide d'un outil de cartographie de corpus également présenté dans cette partie. Enfin, nous abordons la poursuite de nos travaux de recherche et nous concluons sur les résultats de cette étude.

2. PROJET ISOMETA

Les travaux présentés dans cet article s'inscrivent dans la poursuite du projet IsoMeta, initié en 2001, pour mener une analyse des métaphores conceptuelles fondée sur une représentation sémique. L'objet d'étude concerne donc le phénomène autant cognitif que linguistique introduit dans Lakoff et Johnson (1980) sous le nom de métaphore conceptuelle. Le principe central est l'existence d'images récurrentes donnant lieu à de nombreux emplois métaphoriques dans la langue. Ces emplois ne sont pas les seuls, des comparaisons explicites peuvent les accompagner et le caractère conventionnel de nombreuses métaphores conceptuelles étudiées par Lakoff est aussi souvent lié à l'existence d'emplois figés, de catachrèses. Nous renvoyons à Ferrari (2006) pour une présentation plus détaillée de la notion et de travaux qui y sont liés.

Indépendamment de la nature précise des figures de rhétorique sous lesquelles se réalise une métaphore conceptuelle, elle offre avant tout l'intérêt de fournir une connaissance *a priori* exploitable en TAL : l'existence d'un lien privilégié entre le domaine source et le domaine cible. L'exploitation de ce type de lien est désormais récurrente pour coder des connaissances préalables concernant les métaphores conceptuelles conventionnelles : lexique sémantique de grande taille dans Chibout *et al.* (2001), enrichissement de la version italienne d'EuroWordNet par Alonge et Castelli (2003) et Ulivieri *et al.* (2004), renouvellement dans Veale (2003) de l'approche de Gentner (1988) initialement adaptée à l'analogie.

Dans la première phase du projet IsoMeta, nous nous sommes intéressés à une analyse locale de ce type de métaphore conceptuelle afin de dégager quelle dynamique sémique est à l'œuvre dans les emplois lexicaux associés. Nous avons mené une étude sur corpus d'une métaphore conceptuelle conventionnelle spécifique : la « *météorologie boursière* ». Ces travaux ont permis de dégager une caractérisation des sèmes transportés dans les emplois métaphoriques, en relation avec les domaines source et cible de la métaphore conceptuelle sous-jacente (Perlerin *et al.*, 2005).

Dans la poursuite de ces travaux limités à l'étude locale d'emplois métaphoriques, nous cherchons désormais à dégager des régularités à l'échelle d'une collection, en vue de caractériser la dimension intertextuelle des emplois de métaphores conceptuelles. Pour ce faire, nous reprenons le même corpus, constitué d'articles d'économie, présenté en détail dans la section suivante. Nous y étudions conjointement plusieurs métaphores conceptuelles ayant des domaines sources distincts, la *météorologie*, la *santé* et la *guerre*, mais un domaine cible commun : la *bourse*. Le corpus est en effet attesté pour contenir des emplois de ces trois métaphores conceptuelles, de même que des utilisations autres des lexiques des trois domaines sources qui y conservent donc une grande partie de leur polysémie. Les exemples suivant l'illustrent sur deux emplois du mot *orage* :

> *« Paris laisse passer l'**orage** et campe sur des positions relativement solides. »*
>
> *« À la suite du violent **orage** qui avait éclaté dans la nuit de lundi à mardi derniers, les logiciels, pieds dans l'eau, ont refusé de fonctionner. »*

Pour mener une étude à l'échelle d'une collection, nous envisageons d'exploiter des outils d'observation construisant des représentations (visuelles) globales d'un corpus. Ces outils de cartographie documentaire permettent usuellement de regrouper les documents selon des critères de proximité thématique. Dans les expériences que nous présentons ici, il est important de noter que nous les détournons d'emblée de leur usage premier, en utilisant un corpus qui est déjà thématiquement homogène, et en cherchant à y observer des regroupements liés cette fois aux domaines sources des trois métaphores conceptuelles étudiées. Notre objectif principal consiste à mettre en évidence comment s'organisent les emplois de différentes métaphores conceptuelles ayant un même domaine cible, si elles s'excluent les unes les autres ou si elles sont cohérentes au point de permettre d'organiser conjointement le discours. Nous pensons également que des proximités entre textes peuvent nous aider à interpréter le lexique en rapport avec les métaphores conceptuelles que nous étudions, puisqu'il demeure polysémique.

3. OBJETS D'ÉTUDE

3.1. Ressources lexicales

Afin de représenter les domaines sources de la *météorologie*, la *santé* et la *guerre* des métaphores conceptuelles étudiées, nous avons choisi de regrouper des unités lexicales désignant des entités étroitement liées à chacun des trois domaines. Par exemple, le domaine de la météorologie peut être représenté par les unités lexicales suivantes : *soleil*, *pleuvoir*, *tempête*, *temps calme*, *météorologue*, etc.

Pour construire les ensembles de lexies, nous avons utilisé des outils logiciels réalisant différentes analyses fréquentielles (Perlerin, 2002), des coloriages interactifs de lexies au sein de documents du corpus (Beust, 2002) et des mises en contextes de lexies et de leurs

flexions (description de ces outils ainsi que des ressources construites à l'adresse suivante : http://www.info.unicaen.fr/~troy/isometa).

De cette manière, trois ensembles de lexies ont pu être construits, un ensemble par domaine source. La taille des ensembles de lexies varie selon les domaines : 64 lexies représentent le domaine de la guerre, 112 celui de la météorologie et 111 celui de la santé.

Bien qu'intuitive, cette représentation des domaines en ensembles de lexies n'est pas sans poser problème. En effet, la polysémie de certaines lexies reste présente. Par exemple, la lexie *dépression* peut être associée aussi bien au domaine de la météorologie qu'à celui de la santé. Nous avons choisi dans cette étude de placer une telle lexie dans les deux ensembles.

3.2. Corpus

Pour mener nos observations, nous avons constitué un corpus de 303 articles de taille variable (de 200 à 2000 graphies) issus du journal *Le Monde* sur CD-ROM, tous relatifs au domaine de la bourse et totalisant environ 250 000 graphies. Ce corpus, appelé `Corpus_Travail`, couvre la période 1987 – 1989 et contient 46 % de bilans boursiers et 54 % de dépêches portant sur l'actualité boursière des trois années. Les bilans sont reconnaissables par leurs titres contenant soit le mot « `Bilan` », soit une chaîne « `Semaine du` *X* `au` *Y Mois* » où *X*, *Y* et *Mois*, sont le jour du début, le jour de fin et le mois de la période considérée.

Ce corpus est un sous-ensemble du corpus `Corpus_Total` de 594 articles utilisé dans la phase initiale du projet IsoMeta (Perlerin et al., 2005). À partir de ce corpus initial, nous avons sélectionné les articles contenant deux occurrences ou plus de lexies (ou de formes fléchies de ces lexies) d'un ou plusieurs de nos domaines d'étude, soit 303 articles. Par cette sélection, nous faisons l'hypothèse que nous limitons dans les analyses portant sur `Corpus_Travail` le possible « bruit » causé par des articles contenant très peu d'occurrences des domaines étudiés. Les exemples suivants illustrent des emplois de chacune des trois métaphores conceptuelles dans le corpus de travail :

> « Le dénouement dans la **bataille** autour de la première banque commerciale privée du pays a eu peu d'effet sur les cours. » 27/02/1989
>
> « Une véritable **tempête** de hausses, alimentée par une **marée** de capitaux, étrangers pour partie, en quête de placement. » 03/08/1987
>
> « (…) la **plaie** ouverte par le krach est presque **cicatrisée**. » 27/06/1988

La partie suivante présente les résultats d'expériences portant sur la répartition des trois lexiques sur l'ensemble des documents du corpus.

4. EXPÉRIENCES RÉALISÉES

4.1. Préliminaires statistiques

Avant d'observer la répartition des domaines sources des métaphores conceptuelles étudiées, nous avons réalisé une première analyse quantitative mettant en évidence le « passage » du corpus exploité initialement dans le projet IsoMeta (`Corpus_Total`) à notre corpus d'étude actuel (`Corpus_Travail`) :

Domaine	Nb de lexies de la ressource	Nb total de lexies et de leurs flexions :		Nombre d'articles de `Corpus_Total` contenant au moins :	
		`Corpus Total`	`Corpus Travail`	**une lexie du domaine**	**deux lexies du domaine**
Guerre	64	916	831	320 articles → (- 32 %) → 216 articles	
Météo	112	627	569	252 articles → (- 36 %) → 160 articles	
Santé	111	567	485	282 articles → (- 46 %) → 151 articles	

Tableau 1 : Première répartition des ressources lexicales utilisées dans le corpus d'étude

Le tableau 1 met en évidence un nombre plus important de lexies du domaine de la guerre dans les articles malgré la taille réduite de la ressource lexicale utilisée. Il est intéressant de remarquer que des proportions très importantes d'articles ne contiennent qu'une seule occurrence de lexies des domaines d'étude. C'est à cause de ces fortes proportions que nous avons constitué le corpus `Corpus_Travail` afin de favoriser, dans un premier temps, le plus possible la redondance des domaines au sein des articles.

Nous allons maintenant considérer qu'un article de `Corpus_Travail` est étiqueté `Guerre` s'il possède au moins deux occurrences du domaine de la guerre (procédure identique pour les autres domaines). Un article possède donc au moins une étiquette (par constitution de `Corpus_Travail`) et peut en avoir plusieurs s'il contient plus de deux occurrences de plusieurs domaines. La figure 1 met en évidence les éventuels « liens » entre les trois domaines dans les articles ainsi étiquetés. Nous pouvons observer que le domaine de la météorologie est assez lié avec les autres domaines : sur les 160 articles étiquetés `Météo` seulement 28 possèdent cette seule étiquette alors que 73 et 40 articles ne sont respectivement étiquetés que par `Guerre` et `Santé`. Pour aller plus loin dans nos observations, nous proposons dans la partie suivante d'exploiter des représentations cartographiques du corpus afin de naviguer sur ce dernier de manière globale.

4.2. Cartographie du corpus d'étude

Pour construire des représentations cartographiques du corpus d'étude, nous utilisons la plate-forme ProxiDocs. Pour une présentation détaillée de l'outil et un état de l'art sur la cartographie de corpus, nous renvoyons à Roy (2005). Notre objectif à travers cette plate-forme est de fournir des interfaces de lecture rapide d'ensembles documentaires en mettant en place une interaction entre le matériau langagier étudié et l'utilisateur. Nous appelons de telles interfaces de lecture rapide des cartes du corpus. Ces cartes permettent à l'utilisateur de naviguer sur le corpus d'étude et de visualiser de façon interactive différentes informations telles des similarités et des différences entre textes du corpus par rapport aux domaines pris en considération. Afin d'opérationnaliser la construction des cartes du corpus, nous mettons en œuvre dans la plate-forme ProxiDocs différents traitements

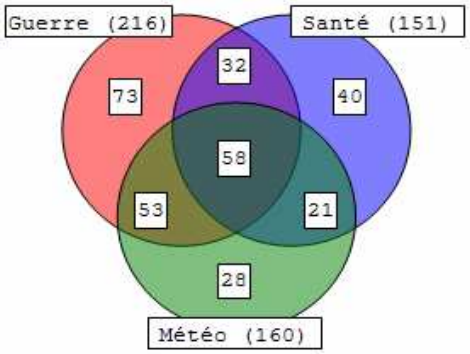


Figure 1 : représentation des ensembles d'articles étiquetés `Guerre`, `Météo` et `Santé` et de leurs intersections, les valeurs sont les cardinalités des intersections et de leurs complémentaires

statistiques (Roy et Beust, 2004). Le premier de ces traitements est un comptage des occurrences de lexies de chaque domaine et de leurs formes fléchies dans chaque texte du corpus. Ce comptage est pondéré selon la taille des textes. Un vecteur de réels de dimension égale au nombre de domaines est alors associé à chaque texte. Afin de visualiser les vecteurs représentant les textes du corpus sur des espaces en 2 dimensions (les cartes), il faut réaliser une étape de projection. Les cartes présentées ci-dessous ont été obtenues à l'aide d'une Analyse en Composantes Principales (Bouroche et Saporta, 1980, p. 17).

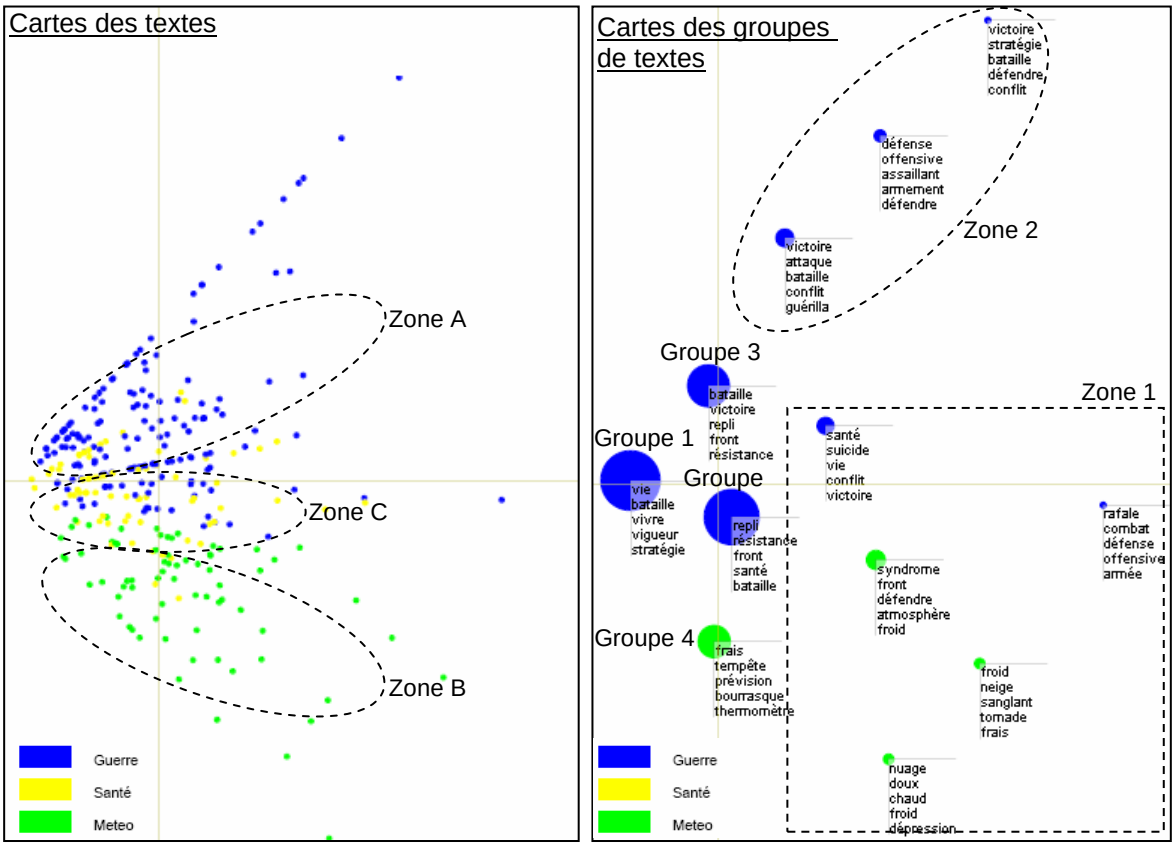


Figure 2 : Cartes du corpus construites par la plate-forme ProxiDocs à partir de nos domaines. Versions SVG de ces cartes : http://www.info.unicaen.fr/~troy/isometa/cartes/

La carte de gauche représente les articles alors que la carte de droite met en évidence des groupes d'articles étiquetés avec les cinq lexies des domaines les plus fréquentes dans les articles qu'ils contiennent. Les couleurs associées aux textes et aux groupes de textes sur les cartes correspondent aux domaines majoritairement représentés. Pour plus de lisibilité, certains groupes et zones ont été marqués manuellement sur les cartes.

La carte des textes révèle que les domaines de la guerre et de la météorologie sont les plus représentés dans les articles du corpus (respectivement en zones A et B). Une lecture des articles de ces deux zones laisse paraître que la zone B (météorologie) contient une majorité de dépêches liées à l'actualité boursière alors que la zone A (guerre) semble contenir autant de bilans que de dépêches. Le domaine de la santé est moins représenté et il est intéressant de voir que les articles contenant majoritairement des lexies de ce domaine se « mélangent » en zone C avec des articles majoritairement du domaine de la guerre. Une lecture rapide des articles de cette zone révèle une proportion importante de bilans.

La carte des groupes de textes met en évidence le résultat d'une Catégorisation Hiérarchique Ascendante (Bouroche et Saporta, 1980, p. 54) de l'espace des documents présenté dans la carte des textes. Cette CHA a mis automatiquement en évidence 12 groupes de textes. La carte des groupes de textes confirme la prédominance des domaines de la guerre et de la météorologie. Sur les 12 groupes d'articles représentés sur la carte, 8 groupes contiennent des articles majoritairement dans le domaine de la guerre, 4 dans le domaine de la météorologie et aucun dans le domaine de la santé. Le tableau suivant résume les propriétés des groupes et des zones marqués sur la carte.

Groupe / Zone	**Nombre d'articles**	**Type d'articles**		**Nombre total de lexies de domaine**		
		Bilans	**Dépêches**	**Guerre**	**Météorologie**	**Santé**
Groupe 1	98	47 %	53 %	162	79	150
Groupe 2	84	73 %	27 %	231	222	185
Groupe 3	49	49 %	51 %	217	37	56
Groupe 4	29	31 %	69 %	18	110	31
Zone 1 (5 groupes)	25	32 %	68 %	90	105	54
Zone 2 (3 groupes)	14	7 %	93 %	101	1	8

Tableau 2 : Propriétés des groupes et des zones marqués sur la carte. Ces informations sont accessibles à partir de la carte des groupes de textes, chaque disque étant un lien hypertexte vers un rapport détaillé sur les articles du groupe représenté.

Nous pouvons alors remarquer que le domaine de la guerre est fortement présent dans les groupes 1 et 3. Les proportions de bilans et de dépêches de ces groupes sont très proches des proportions globales du corpus (soit 46 % de bilans et 54 % de dépêches). Les métaphores conceptuelles remarquées dans les articles de ces deux groupes font intervenir les domaines sources de la guerre et, dans une moindre mesure, de la santé. Toutefois le nombre de ces métaphores est assez faible et les emplois sont plutôt figés.

Le groupe 2 contient une majorité de bilans boursiers. Les métaphores conceptuelles observées font intervenir les trois domaines sources étudiés, leur nombre est plus important que dans les groupes 1 et 3 et les emplois métaphoriques sont plus variés et originaux. Les emplois conjoints des trois métaphores étudiées semblent être propres aux articles de ce groupe et plus particulièrement aux bilans boursiers. Une telle information peut alors nous permettre d'aller vers une caractérisation plus fine de ce type d'articles. Le groupe 4 contient en majorité des dépêches. Un nombre important de métaphores conceptuelles faisant intervenir le domaine source de la météorologie y est observé. Tout comme dans les articles du groupe 2, les emplois métaphoriques sont variés et originaux. Ainsi dans l'exemple du tableau 3 associé à ce groupe, un « *petit vent frisquet* » est l'unique utilisation de « *frisquet* » dans le corpus. La zone 1 marquée sur la carte met en évidence cinq petits groupes de textes (contenant de 1 à 10 articles). Beaucoup de métaphores conceptuelles de domaines sources de la guerre et de la météorologie ont été observées dans les articles de

ces groupes. Au contraire, la zone 2 rassemble trois groupes (contenant de 1 à 9 articles) où un grand nombre de lexies du domaine de la guerre sont présentes. Les emplois des lexies de ce domaine sont non métaphoriques et relèvent des sujets abordés dans ces articles, à savoir, de conflits armés ayant des répercutions sur le marché boursier. Le tableau suivant résume et illustre les différentes observations que nous avons réalisées sur la carte des groupes de textes.

Groupe / Zone	Métaphores présentes	Exemples extraits du corpus
Groupe 1	Figées	« *En neuf mois, six firmes sur les trente-trois OPA ont été l'objet de véritables* ***batailles*** *boursières.* » 26/09/1988
Groupe 2	Variées	« *(…) après avoir* ***contaminé*** *New-York et Londres, la* ***fièvre*** *des OPA s'est mise (…) à ronger la Bourse.* » 08/02/1988
Groupe 3	Figées	« *Selon le SNUI, qui rappelle que le* ***conflit*** *des impôts dure depuis sept mois, (…)* » 22/09/1989
Groupe 4	Variées	« *Un petit* ***vent frisquet*** *a* ***soufflé****, ces derniers jours rue Vivienne, qui, sans crier gare, s'est soudain éclipsé à la dernière minute pour laisser la place à une* ***brise*** *nettement plus* ***chaude****.* » 15/05/1989
Zone 1	Variées	« *Porteur du terrible* ***virus*** *de la défiance, il se propage à la vitesse de l'éclair et les tentatives désespérées de* ***réanimation*** *(…) sont inopérantes.* » 30/10/1987
Zone 2	Aucune	« *Pour se déplacer (..), des* ***officiers*** *de la* ***guérilla*** *utilisent les motos récupérées pendant les* ***attaques****.* » 13/04/1987

Tableau 3 : Illustrations des métaphores présentes dans les groupes et les zones mises en évidence sur la carte de textes. Les extraits du corpus ont été obtenus à partir des rapports d'analyse de chaque groupe accessibles à partir de la carte des groupes de textes.

Il est intéressant de constater la répartition sur la carte des différents groupes en rapport avec les types de métaphores que nous y avons observés (tableau 3) : les emplois littéraux ou les métaphores figées sont situés dans la partie supérieure de la carte, tandis que les emplois variés et plus originaux sont globalement situés dans la partie inférieure. Une gradation de cette répartition peut même être observée si l'on remarque que les emplois figés sont situés entre les emplois littéraux et les métaphores plus vives. L'outil de cartographie exploité semble donc permettre de distinguer les types d'emplois lexicaux associés aux domaines étudiés, permettant ainsi une certaine caractérisation de l'intertextualité.

5. CONCLUSION ET PERSPECTIVES

Nous avons montré dans cet article l'apport d'analyses globales, tant quantitatives que cartographiques, sur un corpus pour y caractériser les usages de trois métaphores conceptuelles. La répartition des lexiques des trois domaines sources a tout d'abord montré des utilisations conjointes des trois métaphores qui ne sont donc pas exclusives l'une de l'autre. Des localisations de certaines métaphores dans des groupes de documents ont également été mises en évidence. La cartographie a permis d'isoler des groupes de textes

caractérisés par des emplois lexicaux d'une nature unique : soit littéraux, soit figés, soit variés. Cette première expérience était fondée sur l'exploitation de simples sacs de mots. Les résultats obtenus nous encouragent à poursuivre nos travaux dans le sens d'une caractérisation plus fine de l'intertextualité.

C'est pourquoi, dans la phase suivante du projet IsoMeta, nous cherchons à reprendre le modèle de structuration différentielle du lexique d'un domaine, exploité dans la première phase du projet, afin de représenter les trois domaines sources étudiés. Ce modèle, intitulé Lucia (Perlerin, 2004), inspiré de la Sémantique Interprétative (Rastier, 1987), permet à un utilisateur d'exprimer des connaissances et un point de vue sur le lexique d'un domaine en organisant des lexies selon deux critères principaux : des regroupements par similarité, témoignant de la proximité de certaines lexies ; et des oppositions locales, précisant les différences entre lexies proches. L'exploitation d'un tel modèle dans notre étude devrait permettre de dégager la dynamique sémique à l'œuvre dans les emplois lexicaux associés aux trois métaphores étudiées à l'échelle du corpus.

Par ailleurs, le corpus sur lequel nous travaillons est constitué d'article traitant de l'actualité boursière d'une période donnée. Le contenu des articles est donc fortement influencé par des événements extérieurs ayant des répercutions sur le monde boursier. Les premières observations ont permis de remarquer un lien assez fort entre la présence de certaines métaphores et l'actualité. Ainsi, la métaphore conceptuelle de la *guerre économique* a été très utilisée lors du mini-krach boursier de fin 1987, alors que celle de la *météorologie boursière* a été très employée lors de conditions climatiques remarquables, hiver rude ou forte chaleur. Il semble également que des utilisations conjointes de métaphores puissent être étroitement liées à l'actualité. Par exemple, la métaphore de la *santé financière* semble particulièrement utilisée lors de grandes crises économiques en complément d'autres métaphores conceptuelles. Nous envisageons d'exploiter à nouveau la cartographie de corpus mais pour mener cette fois une étude diachronique des métaphores conceptuelles précédentes.

Le projet IsoMeta a débuté en 2001 et a donné lieu à de nombreux travaux sur les métaphores conceptuelles. Le tableau suivant résume ces différents travaux :

	Métaphores conceptuelles étudiées	Approche adoptée dans la structuration du lexique des domaines sources	Corpus étudié
Phase 1 (Perlerin *et al.*, 2005)	*Météorologie boursière*	Approche Sémantique Différentielle	Corpus Total
	Dégager localement la dynamique sémique à l'œuvre dans les emplois lexicaux associés aux métaphores conceptuelles		
Phase 2 (travaux actuels)	*Météorologie boursière*, *Guerre économique* et *Santé financière*	Approche « sacs de mots »	Corpus Travail
	Analyser en corpus des trois métaphores conceptuelles afin d'observer leurs interactions et d'aider à interpréter le lexique correspondant		
Phase 3 (travaux à venir)	*Météorologie boursière*, *Guerre économique* et *Santé financière*	Approche Sémantique Différentielle	Corpus Total
	Analyser à l'échelle du corpus la dynamique sémique à l'œuvre dans les emplois lexicaux associés aux métaphores conceptuelles, mettre en évidence l'influence de l'actualité dans ces emplois		

Tableau 4 : Les différentes phases du projet IsoMeta

L'un des objectifs de ce projet est de contribuer à une problématique de sémantique lexicale qui dépasse l'étude des métaphores. À travers nos travaux, nous cherchons à apporter notre contribution à la linguistique de corpus, en particulier en proposant une certaine forme d'instrumentalisation avec la plateforme ProxiDocs. Cet outil permet, dans la constitution d'un corpus comme dans ses analyses, les études en synchronie et en diachronie, et répond en partie au besoin d'analyses multi-échelles (paragraphe, document, groupe de documents, corpus dans son ensemble par exemple). L'enjeu de ce type d'instrumentalisation est de mieux appréhender les rapports du local au global dans la construction du sens.

6. REFERENCES

Alonge A., Castelli M. 2003. 'Encoding Information on Metaphoric Expressions in Word-Net-like Resources', *ACL 2003 Workshop on the Lexicon and Figurative Language*, 10-17.

Beust P. 2002. 'Un outil de coloriage de corpus pour la représentation de thèmes', *Actes des Journées internationales de l'Analyse des Données Textuelles*.

Bouroche J.-M., Saporta G. 1980. *L'Analyse des Données*, Paris : Presses Universitaires de France.

Chibout K., Vilnat A., Briffault X. 2001. 'Sémantique du lexique verbal : un modèle en arborescence avec les graphes conceptuels', *TAL, Lexiques sémantiques*, Vol. 42-3, p. 691-727.

Ferrari, S. 2006. 'Rhétorique et compréhension', in Sabah, G. (dir), *Compréhension des langues et interaction*, chapitre 6. Traité IC2, Hermès Science, Lavoisier.

Gentner D. 1988. 'Analogical Inference and Analogical Access', *Analogica*, chap. 3, London : A. Prieditis (ed), Pitman Publishing, Morgan Kaufmann Publishers, p. 63-88.

Lakoff G., Johnson M. 1980. *Metaphors we live by*. Chicago : University of Chicago Press.

Perlerin V. 2004. *Sémantique légère pour le document : assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Thèse d'informatique de l'Université de Caen.

Perlerin V., Ferrari S., Beust P. 2005. 'Métaphores et dynamique sémique : expériences sur corpus', in Williams G. (dir), *La Linguistique de Corpus*, Rennes : Presses Universitaires de Rennes, p. 323-336.

Perlerin V. 2002. 'Memlabor, un environnement de création, de gestion et de manipulation de corpus de textes', *Actes des Rencontres des Étudiants Chercheurs en Informatique et en Traitement Automatique des Langues* dans le cadre de la conférence *TALN*, p. 507-516.

Rastier F. 1987. *Sémantique Interprétative*, Paris : Presses Universitaires de France.

Roy T. 2005. 'Une plate-forme logicielle dédiée à la cartographie thématique de corpus', *Actes des Rencontres des Étudiants Chercheurs en Informatique et en Traitement Automatique des Langues* dans le cadre de la conférence *TALN*, p. 545-554.

Roy T., Beust P. 2004. 'ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus', *Actes des Journées internationales de l'Analyse des Données Textuelles*, p. 978-987.

Ulivieri M., Guazzini E., Bertagna F., Calzolari N. 2004. 'Senseval-3 : The Italian All-words Task', SENSEVAL-3 : *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Association for Computational Linguistics.

Veale T. 2003. 'Systematicity and the Lexicon in Creative Metaphor', *ACL 2003 Workshop on the Lexicon and Figurative Language*, p. 28-35.

LES CORPUS SPECIALISES AU SERVICE DE LA SEMANTIQUE QUANTITATIVE : LA POLYSEMIE DU FRANÇAIS TECHNIQUE

Ann Bertels
ILT – K.U.Leuven / QLVL – K.U.Leuven (Belgique)
Dirk Speelman, Dirk Geeraerts
QLVL – K.U.Leuven (Belgique)

RÉSUMÉ

Cet article décrit la méthodologie pour procéder à une analyse sémantique quantitative dans un corpus spécialisé, en l'occurrence un corpus relevant du domaine technique des machines-outils pour l'usinage des métaux. À partir du contexte linguistique, cette étude vise à vérifier si et dans quelle mesure les mots spécifiques du corpus technique (ou corpus de langue spécialisée) sont monosémiques ou polysémiques. À cet effet, le corpus technique sera soumis à des exploitations statistiques pour identifier les mots spécifiques ou spécificités ainsi qu'à des analyses de cooccurrences permettant de déterminer le degré de monosémie de ces spécificités. Les données quantitatives feront également l'objet d'analyses statistiques, dont nous présenterons les premiers résultats.

MOTS-CLES

Corpus spécialisé, sémantique quantitative, spécificités, polysémie, cooccurrences.

1. INTRODUCTION ET QUESTION DE RECHERCHE

Cet article s'inscrit dans le cadre d'une étude sémantique quantitative du vocabulaire spécifique d'un corpus en français technique. Le but de cette étude est d'étudier la sémantique (la monosémie ou la polysémie) des mots et termes spécifiques du corpus technique. Les textes de ce corpus spécialisé relèvent tous du domaine technique des machines-outils pour l'usinage des métaux. Les corpus électroniques et surtout les corpus de textes spécialisés sont très précieux et même indispensables pour l'analyse sémantique, car ils permettent d'accéder au contexte linguistique. Les corpus électroniques ainsi que les outils informatiques de linguistique de corpus facilitent les analyses linguistiques et statistiques. Toutefois, l'exploitation de ces grandes quantités de textes requiert une approche quantitative et automatisée et donc un accès automatisé au contexte linguistique, surtout si l'on procède à une analyse sémantique de quelque 5000 mots. L'objet de cet article est de montrer qu'une analyse sémantique quantitative et automatisée est possible et qu'elle permet de quantifier la monosémie ou la polysémie à partir d'analyses de corpus, plus particulièrement à partir d'analyses de cooccurrences de cooccurrences.

Pour étudier la monosémie ou la polysémie dans un corpus technique, nous proposons d'articuler la question de recherche autour de deux axes méthodologiques : premièrement, l'identification du vocabulaire spécifique, c'est-à-dire les spécificités[(1)], et deuxièmement,

la détermination de la monosémie ou polysémie de ces spécificités. À partir du degré de spécificité et du degré de monosémie, les spécificités seront situées sur un continuum de spécificité ainsi que sur un continuum de monosémie. Par conséquent, la question de recherche principale étudie la corrélation entre le continuum de spécificité et le continuum de monosémie, par l'intermédiaire d'une analyse de régression simple. L'hypothèse de recherche avancée pose que les mots (les) plus spécifiques du corpus technique ne sont pas nécessairement (les) plus monosémiques, au contraire même. En effet, certains mots spécifiques du corpus technique sont des mots à sens multiples. Citons par exemple le mot *broche* (1) « partie tournante d'une machine-outil qui porte un outil ou une pièce à usiner » et (2) « outil servant à usiner des pièces métalliques ». Signalons également le mot *découpe* (1) « action de découper » et (2) « résultat de la découpe (pièce découpée) », où les deux sens indiquent clairement une relation métonymique. Comme les spécificités sont situées sur un continuum de spécificité et sur un continuum de monosémie, l'approche adoptée est une approche scalaire par continuum. Cette approche par continuum rend la question de recherche mesurable et objectivable et permet ainsi de l'implémenter pour un nombre important de mots et de contextes linguistiques. Cependant, le degré de monosémie n'est pas uniquement influencé par le degré de spécificité, mais également par d'autres facteurs comme la fréquence dans le corpus technique, la fréquence dans le corpus de référence de langue générale, la classe lexicale, le nombre de classes lexicales et la longueur. Dès lors, il est important de faire intervenir dans les analyses statistiques de régression tous les facteurs influant sur le degré de monosémie. Une analyse de régression multiple permettra d'intégrer tous les facteurs déterminants (c'est-à-dire toutes les variables indépendantes) et d'évaluer leur impact combiné sur le degré de monosémie.

Cette étude sémantique quantitative s'articule autour de deux axes méthodologiques et vérifie la corrélation entre le continuum de spécificité et le continuum de monosémie dans un corpus technique. Le corpus technique est constitué de textes spécialisés (environ 1,7 million d'occurrences), relevant du domaine technique des machines-outils pour l'usinage des métaux. Il a été étiqueté par Cordial 7 Analyseur et consiste en 4 sous-corpus, datant de 1998 à 2001 : revues électroniques (800.000) et fiches techniques (300.000) trouvées sur Internet, normes ISO et directives (300.000) et quatre manuels numérisés (360.000). Les textes se situent à différents niveaux de normalisation et de vulgarisation, ce qui assure la représentativité du corpus technique spécialisé. Afin de pouvoir déterminer les spécificités, le corpus technique est complété par un corpus de référence de langue générale. Le corpus de référence est constitué d'articles du journal *Le Monde* (janv.–sept.1998). Il a aussi été étiqueté et comprend environ 15.300.000 mots. Le corpus technique et le corpus de référence respectent donc le rapport de 1 à 10. Les fichiers générés par Cordial 7 Analyseur se composent de trois colonnes, à savoir (1) la forme graphique, (2) le lemme et (3) le code Cordial, comparable à un POS-tag et indiquant la classe lexicale.

Dans cet article, nous présenterons une double méthodologie quantitative pour accéder aux informations sémantiques renfermées dans le corpus spécialisé. Après la justification méthodologique (section 2), nous expliquerons l'approche quantitative, d'une part au niveau des spécificités (section 3) et, d'autre part, au niveau de l'analyse sémantique de ces spécificités (section 4). Ensuite, les données quantitatives de spécificité et de monosémie (ou polysémie) feront l'objet d'une analyse statistique de régression simple, dont nous présenterons les premiers résultats (section 5). Par conséquent, notre étude quantitative se situe à l'entrecroisement de trois disciplines : la linguistique de corpus, l'informatique et la statistique.

2. JUSTIFICATION METHODOLOGIQUE

Pour étudier la sémantique dans le corpus technique, nous adoptons une approche quantitative et une approche scalaire par continuum. Toutefois, pour étudier la sémantique dans la langue spécialisée, l'approche adoptée traditionnellement est une approche catégorielle. Cette approche catégorielle se caractérise, d'une part, par un classement catégoriel ou binaire des unités linguistiques, c'est-à-dire par la dichotomie *termes – mots*. D'autre part, elle se caractérise par un classement catégoriel ou binaire au niveau sémantique, à savoir l'opposition *monosémie – polysémie*. La terminologie traditionnelle préconise la monosémie et l'univocité des termes d'une langue de spécialité, tandis que polysémie est réservée aux mots de la langue générale.

Comment justifier le choix d'une approche scalaire par continuum pour cette étude sémantique quantitative, plutôt qu'une approche catégorielle traditionnelle ? Récemment, la linguistique de corpus ainsi que l'analyse de corpus spécialisés ont remis en question l'approche onomasiologique prescriptive de la terminologie traditionnelle et contribué à l'essor de la terminologie descriptive. Sur le plan des unités linguistiques, les partisans de la terminologie descriptive rejettent la dichotomie entre langue générale et langue spécialisée et adoptent une approche sémasiologique et linguistique (Condamines & Rebeyrolle, 1997). Les termes font partie intégrante de la langue naturelle, mais ils se caractérisent par le fait qu'ils véhiculent des connaissances spécialisées (Lerat, 1995). En plus, le vocabulaire d'un corpus technique ne contient pas uniquement des mots techniques ou « termes » au sens strict, qui sont propres au domaine de spécialité, tels que *usinage* ou *broche*, mais également des mots du VGOS (vocabulaire général d'orientation scientifique) (Phal, 1971). Ces mots s'emploient dans plusieurs domaines scientifiques et techniques et leur sens est déterminé par les contextes spécialisés (*machine*, *outil*). Finalement, le vocabulaire d'un corpus technique comprend également des mots de la langue générale, tels que *type*, *modèle*, *permettre*, etc. Sur le plan de l'analyse sémantique, la monosémie et l'univocité de la langue spécialisée ont été remises en question notamment par la Théorie Communicative de la Terminologie (Cabré, 1998 et 2000), par la Socioterminologie (Gaudin, 1993) et par la Terminologie socio-cognitive (Temmerman, 1997 et 2000). Ces remises en questions nous incitent donc à adopter une approche par continuum, aussi bien au niveau des unités linguistiques (continuum de spécificité) qu'au niveau sémantique (continuum de sens ou continuum de monosémie).

3. IDENTIFIER LES SPECIFICITES

Le premier axe méthodologique vise à identifier les spécificités du corpus technique. Les spécificités d'un corpus ne sont pas les mots les plus fréquents, mais les mots les plus caractéristiques et les plus représentatifs du corpus, c'est-à-dire les mots qui caractérisent thématiquement le corpus. Les spécificités figurent, d'un point de vue relatif, de façon significative plus fréquemment dans le corpus technique spécialisé que dans un corpus de référence de langue générale. À cet effet, les fréquences dans le corpus technique spécialisé (corpus d'analyse) sont comparées aux fréquences dans le corpus de référence de langue générale, compte tenu de la taille des deux corpus. Si la fréquence observée dans le corpus d'analyse est supérieure à la fréquence attendue dans le corpus de référence et si cette différence est statistiquement significative, les mots sont spécifiques du corpus technique (spécificités positives).

Afin d'identifier les spécificités du corpus technique, nous utilisons la *Keywords Method*, implémentée dans le logiciel Abundantia Verborum Frequency List Tool[(2)]. Cette méthode se sert de la statistique de test du ratio du log de vraisemblance (*log likelihood ratio* - LLR) (Dunning, 1993). D'autres logiciels et méthodologies sont également disponibles, notamment l'outil *Keywords* du logiciel WordSmith Tools[(3)] pour la méthode des mots-clés. La deuxième méthodologie du calcul des spécificités est basée sur la statistique de la distribution hypergéométrique (Lafon, 1984) et elle est implémentée dans le logiciel Lexico3[(4)]. Le calcul des spécificités procède par comparaison *partie-tout* et compare ainsi une section d'un corpus au corpus entier afin d'identifier le vocabulaire spécifique de cette section. La fréquence relative d'un mot dans la section est donc comparée à la fréquence relative dans le corpus entier. Si la fréquence relative observée (dans la section) est supérieure à la fréquence attendue, c'est-à-dire la fréquence relative dans le corpus entier, il s'agit d'une spécificité positive de la section analysée. L'analyse statistique sous-jacente au calcul des spécificités utilise le test de Fisher Exact, basé sur les probabilités exactes de la distribution hypergéométrique. Les spécificités positives indiquent un suremploi dans la section analysée, tandis que les spécificités négatives signalent un sous-emploi.

Les deux méthodologies et les trois outils mènent à des résultats similaires, à savoir une liste de mots-clés ou spécificités du corpus technique, étant considéré soit comme corpus spécialisé par rapport au corpus de référence général (méthode des mots-clés), soit comme section intégrée dans le corpus de référence pour la comparaison *partie-tout* (méthode du calcul des spécificités). Nous recensons environ 13.000 spécificités positives dans le corpus technique ($p \leq 0.05$). À l'aide des codes Cordial indiquant la classe lexicale, il est possible de dresser une liste de mots grammaticaux (ou mots fonctionnels) (450) et une liste de noms propres (7200), permettant de filtrer la liste de mots-clés. Après suppression des mots grammaticaux, des noms propres ainsi que des hapax (2112), la liste de mots-clés contient environ 5000 spécificités du corpus technique. Pour chaque spécificité, le logiciel Abundantia Verborum Frequency List Tool fournit la fréquence technique (absolue et relative), la fréquence générale (absolue et relative), la mesure statistique du log de vraisemblance (LLR) et le complément de la valeur p (pour $1\text{-}p \geq 0.95$). La mesure statistique indiquant le degré de spécificité permet de classer les spécificités par ordre de spécificité décroissante et d'attribuer un rang de spécificité. Les spécificités avec le même degré de spécificité (LLR) se voient attribuer le même rang de spécificité, le rang 1 indiquant le mot le plus spécifique du corpus technique. Les rangs de spécificité permettent finalement de situer les spécificités sur une échelle ou un continuum de spécificité.

4. QUANTIFIER LA POLYSEMIE À PARTIR DU CONTEXTE

Les analyses de corpus ne permettent pas uniquement d'identifier les mots les plus représentatifs d'un corpus spécialisé par rapport à un corpus de référence, mais elles permettent également d'étudier la sémantique de ces spécificités. Comme il s'agit de l'analyse sémantique de quelque 5000 spécificités, il vaut mieux automatiser l'analyse et essayer de quantifier la sémantique. Afin de réaliser cette analyse sémantique automatisée et quantitative, nous proposons d'implémenter la monosémie comme l'homogénéité sémantique, ce qui constitue une tentative d'opérationnalisation du degré de monosémie. Un mot est d'autant plus monosémique que ses cooccurrents sont sémantiquement homogènes, car les cooccurrents sont des critères désambiguïsateurs puissants (Schütze

1998 ; Véronis 2003). Le caractère monosémique ou polysémique d'un mot de base se caractérise donc par des contextes (ou cooccurrents) sémantiquement homogènes si le mot est monosémique, ou sémantiquement hétérogènes si le mot est polysémique.

Par exemple pour le mot *tour*, nous recensons les cooccurrences suivantes :

> *(a) L'atelier de sous-traitance le fabrique en quatre étapes. La première est un tour sur lequel on usine l'alésage central. Ensuite, un centre d'usinage …*
>
> *(b) La technologie actuelle de la grande vitesse fait appel à …, ce qui convient à des broches pouvant monter jusqu'à quinze mille tours par minute, …*

La première cooccurrence (a) indique clairement le sens « machine-outil pour l'usinage des pièces », tandis que la deuxième cooccurrence (b) montre le mot *tour* dans le sens de « rotation, révolution ».

Toutefois, pour avoir accès à la sémantique des cooccurrences, il convient d'analyser leurs cooccurrents, c'est-à-dire les cooccurrents de deuxième ordre, et plus particulièrement leur recouvrement. En effet, le degré de recouvrement des cooccurrences de deuxième ordre est un indice important du degré de monosémie d'un mot de base, car il indique à quel point les cooccurrences de premier ordre sont sémantiquement homogènes. Si les cooccurrents des cooccurrents (ou « cc ») sont formellement très différents et se recouvrent très peu, les cooccurrents (ou « c ») seront sémantiquement plus diversifiés. Les cooccurrents sémantiquement diversifiés appartenant à plusieurs champs sémantiques, la spécificité aura moins de chances d'être monosémique. En revanche, si les cooccurrents des cooccurrents se caractérisent par un recouvrement (très) important, cela veut dire qu'un nombre (très) important des cooccurrents apparaissent avec le(s) même(s) cc et donc que les cooccurrents sont sémantiquement homogènes. Par conséquent, la spécificité sera plus monosémique et aura un degré de monosémie plus élevé.

Pour déterminer le degré de monosémie, nous proposons une formule (*Cf.* Figure 1) basée sur le recouvrement des cooccurrents des cooccurrents (cc), tenant compte (1) de la fréquence d'un cc dans la liste des cc (= nombre de cooccurrents (c) apparaissant avec ce (cc)), (2) du nombre total de c et (3) du nombre total de cc[5].

$$\boxed{\sum_{\text{cc}} \frac{\text{fq}\, cc}{\#\,\text{total}\, c \cdot \#\,\text{total}\, cc}}$$

Figure 1 : Formule de recouvrement des cooccurrents des cooccurrents

Un cc sera d'autant plus important pour le recouvrement qu'il figure plus souvent dans la liste des cc (si plus de c apparaissent avec ce cc). La somme des poids de tous les cc se situe toujours entre 0 et 1, vu qu'on divise par le nombre de cc. Un poids plus près de 1 indique que les cc sont fortement partagés par les c et que le mot de base sera sémantiquement homogène. Par contre, un poids plus près de 0 indique un faible recouvrement des cc et le mot de base sera sémantiquement plus hétérogène.

La mesure d'association utilisée pour déterminer les cooccurrences statistiquement significatives est la statistique LLR (log de vraisemblance), pour une valeur p très sévère ($p \leq 0.0001$), afin de relever uniquement les cooccurrences sémantiquement pertinentes. Des scripts Python[6] permettent de définir les paramètres (fenêtre d'observation, seuil de significativité, etc.) et génèrent une grande base de données avec toutes les informations statistiques des cooccurrences. Pour toutes les spécificités du corpus technique, nous calculons ainsi le degré de recouvrement et par conséquent le degré de monosémie. Par

analogie avec le rang de spécificité, les spécificités ayant le même degré de monosémie auront également le même rang de monosémie, permettant de situer les spécificités sur une échelle ou un continuum de monosémie.

5. PREMIERS RESULTATS

5.1. Analyse de régression simple

L'étude étant en cours, nous discuterons les premiers résultats de l'analyse statistique. La question de recherche principale étudie la corrélation entre le continuum de spécificité et le continuum de monosémie. À cet effet, on procède à une analyse de régression simple dans le logiciel statistique R[7], avec le rang de monosémie comme variable dépendante (ou expliquée) et le rang de spécificité comme variable indépendante (ou explicative). Le but est d'étudier l'impact de la variable indépendante sur la variable dépendante, c'est-à-dire l'impact du rang de spécificité sur le rang de monosémie. L'analyse de régression simple montre une corrélation négative de -0.71 (coefficient de corrélation Pearson) : plus les mots sont spécifiques, moins ils sont monosémiques. L'analyse est hautement significative (pour $p < 2.2e-16$) et le pourcentage de variation expliquée R^2 est de 51.57 %. La variation du rang de spécificité permet donc d'expliquer 51.57 % de la variation du rang de monosémie.

La visualisation ci-dessous (*Cf.* Figure 2) montre la droite de régression descendante et donc le rapport négatif entre le rang de spécificité et le rang de monosémie. Les mots les plus monosémiques, à gauche de la visualisation, sont clairement les plus polysémiques, car ils se situent en haut (rang de monosémie > 4000). Cette tendance linéaire négative s'oppose au rapport linéaire positif qu'il y aurait selon les monosémistes traditionnels. La terminologie traditionnelle préconise la monosémie des mots techniques de la langue spécialisée et elle réserve la polysémie aux mots de la langue générale.

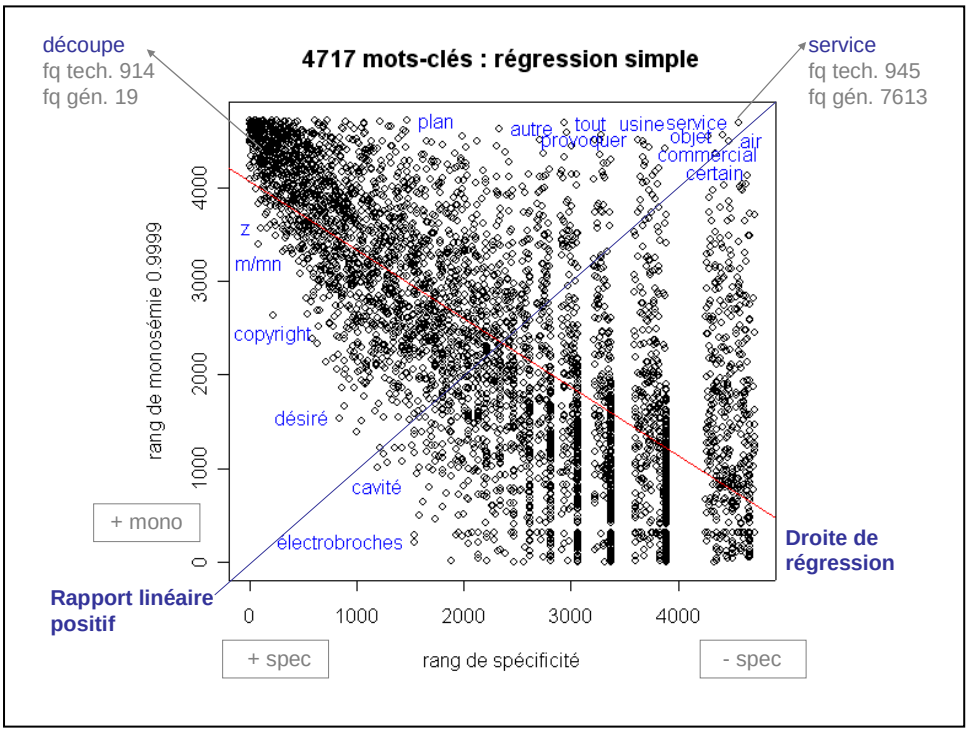


Figure 2 : Visualisation de l'analyse de régression simple

Les résultats de l'analyse de régression simple ainsi que la visualisation ci-dessus indiquent donc que les mots les plus spécifiques de notre corpus technique ne sont pas les plus monosémiques, mais au contraire, les plus polysémiques. En plus, les mots les moins spécifiques sont plutôt monosémiques, à quelques exceptions près, notamment *service, objet, commercial, air*, etc.., qui se situent en haut à droite (*Cf.* Figure 2). Ces exceptions à la tendance générale de corrélation négative sont des mots peu (ou très peu) spécifiques, mais polysémiques. Parmi les mots polysémiques, nous recensons donc les mots les plus spécifiques (*machine, outil, …*), des mots moins spécifiques, ainsi que les mots les moins spécifiques (*service, objet*) du corpus technique. On y retrouve (1) des « termes » au sens strict ou des mots propres au domaine de spécialité, tels que *découpe* et *usinage*, et (2) le « VGOS » ou vocabulaire général d'orientation scientifique, qui s'emploie dans plusieurs domaines scientifiques et techniques avec des mots tels que *machine* et *outil*. Les mots les plus polysémiques comprennent aussi (3) des mots que l'on pourrait qualifier de « mots de la langue générale » (*objet, air*), des mots peu spécifiques dans notre corpus technique.

Ces mots peu spécifiques et polysémiques ont des résidus importants, ce qui signifie que les distances entre les valeurs observées et les valeurs estimées sont très grandes. Si les résidus sont positifs, les mots se situent au-dessus de la droite de régression, par exemple *service*, et ils sont plus polysémiques qu'attendu en fonction de leur rang de spécificité. Si, par contre, les résidus sont négatifs, les mots se situent en dessous de la droite de régression, par exemple *électrobroches*, et ils sont plus monosémiques qu'attendu en fonction de leur rang de spécificité. Comment pourrait-on interpréter la position de ces mots à résidus importants ? Quelle est la raison pour laquelle ces mots se situent tellement loin de la droite de régression ? Ou pourquoi l'analyse de régression simple ne permet-elle pas de bien prédire leur rang de monosémie en fonction de leur rang de spécificité ?

5.2. Explication linguistique

Les mots à résidus positifs sont peu spécifiques et polysémiques et ils se caractérisent par une fréquence technique importante, mais par une fréquence générale beaucoup plus importante en termes relatifs. À titre de comparaison, la fréquence technique de *service* est de 945 et sa fréquence générale de 7613, tandis qu'un mot en haut à gauche polysémique, tel que *découpe*, a une fréquence technique comparable de 914, pour une fréquence générale de 19 (Cf. Figure 2). Les mots à résidus positifs importants, comme *service*, sont des mots généraux : ils sont très peu spécifiques en raison de leur fréquence générale importante. Toutefois, ils sont polysémiques en raison de leur nombre important de c et cc (et en raison de leur fréquence technique importante). Les mots à résidus négatifs importants, tels que *électrobroches*, sont plutôt spécifiques en raison de leur fréquence générale limitée ou zéro et ils sont plus monosémiques qu'attendu en raison du très faible nombre de cc différents, et cela en dépit de leur fréquence technique considérable.

Comme la visualisation de l'analyse de régression simple montre, beaucoup de mots à résidus importants et comme ces résidus positifs et négatifs ne sont pas dispersés de façon équilibrée des deux côtés de la droite de régression, on se voit confronté au problème de l'hétéroscédasticité des résidus. L'hétéroscédasticité signifie que les résidus ne sont pas répartis selon une distribution normale et le problème est détecté à l'aide du test statistique de Goldfeld-Quandt, implémenté dans R. Les solutions techniques adoptées généralement consistent soit en des transformations polynomiales ou logarithmiques des variables de l'analyse de régression simple, soit en une analyse de régression simple pondérée. Les transformations logarithmiques et polynomiales ne permettent pas de résoudre le problème de façon convaincante. La deuxième solution technique, à savoir l'analyse de régression pondérée, est basée sur la méthode des moindres carrés pondérés. La pondération revient à accorder, dans le modèle de régression pondérée, moins d'importance aux mots avec des résidus importants et plus d'importance aux mots avec des résidus limités. Le résultat de la régression pondérée est un bon pourcentage de variation expliquée R^2 de 62,51 % et une tendance linéaire négative plus claire, visualisée ci-dessous (Cf. Figure 3). Les mots sont moins dispersés (ce qui explique le pourcentage R^2 plus élevé), mais ils sont plus comprimés. Le mot *service* se retrouve au même niveau de spécificité, mais plus bas, donc à un niveau plus monosémique. Par conséquent, la régression pondérée ne tient pas compte de l'hétérogénéité sémantique du mot *service*, due à son nombre important de c et de cc et à sa fréquence technique élevée. En plus, un mot plutôt spécifique mais monosémique, tel que *électrobroches*, se retrouve plus à droite et serait donc moins spécifique, ce qui revient à négliger sa fréquence générale limitée ou zéro et à sous-estimer son importance en matière de fréquence technique.

En guise de conclusion, l'analyse de régression pondérée est une bonne solution technique pour le problème technique de l'hétéroscédasticité des résidus. Toutefois, force est de constater que l'analyse de régression pondérée revient à sous-estimer ou méconnaître les caractéristiques des mots (fréquence technique ou générale). Il convient de s'interroger sur la présence ou l'influence d'un facteur caché, qui pourrait nous aider à expliquer pourquoi ces mots à résidus importants se situent aussi loin de la droite de régression et donc pourquoi le modèle de régression ne permet pas de prédire leur rang de monosémie en fonction de leur rang de spécificité. Ainsi, il serait possible de résoudre le problème de l'hétéroscédasticité, tout en respectant les caractéristiques linguistiques. À cet effet, nous élaborons actuellement une mesure de monosémie technique pondérée qui tient compte de la spécificité des cc pour le calcul du recouvrement des cc.

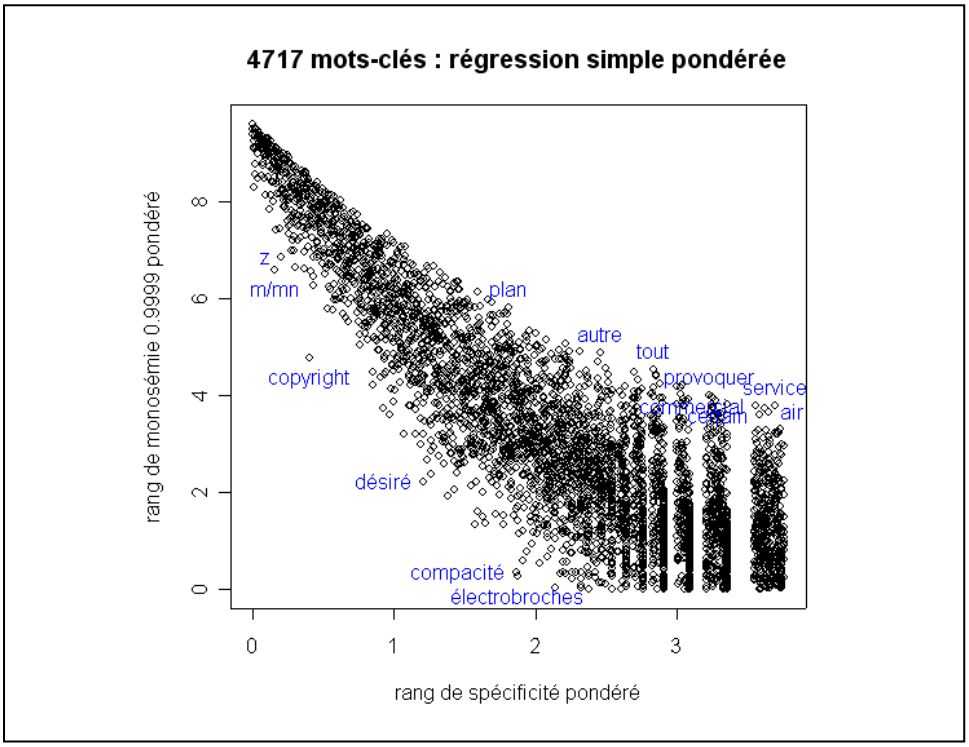


Figure 3 : Visualisation de l'analyse de régression simple pondérée

6. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons présenté la méthodologie ainsi que les premiers résultats d'une étude sémantique quantitative menée sur un corpus spécialisé. Les outils de la linguistique de corpus, plus particulièrement la méthode des mots-clés (*Keywords Method*) et l'analyse des cooccurrences des cooccurrences, ont permis de déterminer d'une part les spécificités du corpus technique et d'autre part leur degré de monosémie, par le biais d'une mesure de recouvrement des cooccurrences des cooccurrences. L'analyse statistique des données quantitatives a démontré et visualisé une corrélation négative entre le rang de spécificité et le rang de monosémie des 5000 spécificités. Les mots les plus spécifiques du corpus technique sont donc les plus polysémiques, tandis que les mots les moins spécifiques sont plutôt monosémiques, à quelques exceptions près, à savoir les mots à résidus importants. Pour ces mots, des recherches supplémentaires s'imposent, en termes de degré de monosémie technique pondérée.

Nous envisageons cette double analyse du degré de spécificité et du degré de monosémie pour les collocations spécifiques également, parce que les termes se situent souvent au niveau des unités polylexicales. Nous nous proposons aussi de procéder à une validation manuelle de la formule déterminant le degré de monosémie à l'aide de l'analyse des cooccurrences relevées. Comme le caractère monosémique ou polysémique d'une unité linguistique n'est pas uniquement influencé par sa spécificité, il est important d'analyser également l'influence d'autres facteurs influant sur le degré de monosémie, notamment la fréquence technique et générale, la classe lexicale, le nombre de classes lexicales, etc. Une analyse de régression multiple permettra d'évaluer l'impact combiné des toutes les variables indépendantes (VI) sur le rang de monosémie (VD).

Nous nous proposons finalement d'effectuer des analyses de régression simple et multiple détaillées, c'est-à-dire par classe lexicale (substantifs / adjectifs / verbes / adverbes) et par sous-catégorie (par exemple les substantifs déverbaux, les abréviations et sigles), ainsi que dans les différents sous-corpus (revues électroniques / fiches techniques / normes et directives / manuels).

7. REFERENCES

Cabré M.T. 1998. *La terminologie. Théorie, méthode et applications*, Ottawa : Les Presses de l'Université.

Cabré M.T. 2000. 'Terminologie et linguistique : la théorie des portes' *Terminologies nouvelles*, 21: 10-15.

Condamines A., Rebeyrolle J. 1997. 'Point de vue en langue spécialisée' *Meta*, 42-1, p. 174-184.

Dunning T. 1993. 'Accurate methods for the statistics of surprise and coincidence' *Computational Linguistics*, 19-1, p. 61-74.

Gaudin F. 1993. *Pour une socioterminologie. Des problèmes sémantiques aux pratiques institutionnelles.* Rouen : Publications de l'Université de Rouen.

Lerat P. 1995. *Les langues spécialisées.* Paris : PUF.

Phal A. 1971. *Vocabulaire général d'orientation scientifique (V.G.O.S.). Part du lexique commun dans l'expression scientifique.* Paris : Crédif – Didier.

Schütze H. 1998. 'Automatic Word Sense Discrimination' *Computational Linguistics*, 24-1, p. 97-123.

Temmerman R. 1997. 'Questioning the univocity ideal. The difference between socio-cognitive Terminology and traditional Terminology' *Hermes* 18, p. 51-90.

Temmerman R. 2000. *Towards new ways of terminology description. The sociocognitive approach.* Amsterdam / Philadelphia : John Benjamins Publishing Company.

Véronis J. 2003. 'Cartographie lexicale pour la recherche d'informations' *Actes de TALN 2003*, p. 265-274.

8. NOTES

1. Nous adoptons le terme « spécificités » pour désigner les mots les plus spécifiques et les plus caractéristiques du corpus technique d'analyse, indépendamment de la méthode utilisée (calcul des spécificités vs. *KeyWords Method*).
2. Abundantia Verborum : http://wwwling.arts.kuleuven.be/genling/abundant
3. WordSmith Tools version 3 : http://www.lexically.net/wordsmith/ et http://www.oup.com
4. Lexico3 : SYLED – CLA2T, Paris3 : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/
5. les doublons comptant double
6. Python : http://www.python.org/
7. logiciel statistique R : http://www.r-project.org/

ANALYSE DISTRIBUTIONNELLE DE CORPUS DE LANGUE GENERALE ET SYNONYMIE

Edith Galy, Didier Bourigault
Equipe de Recherche en Syntaxe et Sémantique
(Université Toulouse le Mirail – CNRS)

1. INTRODUCTION

L'analyse distributionnelle automatique est une méthode qui se fonde sur l'hypothèse inspirée de Z. H. Harris, selon laquelle à partir d'un corpus analysé syntaxiquement on peut rapprocher des couples d'unités sur la base de contextes syntaxiques identiques et induire de cette façon des classes sémantiques de mots (Harris, 1968). En Traitement Automatique des Langues, l'utilisation de techniques d'analyse distributionnelle automatique a été largement explorée sur des corpus spécialisés (Sager, 1987 ; Grefenstette 1994). Parce que les résultats de ce type d'analyse reflètent les usages spécifiques propres aux corpus, ils sont utilisés pour des tâches de construction de ressources terminologiques (Habert & Nazarenko, 1996 ; Bourigault, 2002 ; Bourigault, Aussenac & Charlet, 2004).

L'idée que nous proposons aujourd'hui est de quitter les corpus de langue spécialisée pour appliquer l'analyse distributionnelle automatique sur des corpus de « langue générale », *« language as a whole »* selon Harris (1988). Mais officier sur de tels corpus est un pari encore audacieux : les réserves sont nombreuses. Harris, le premier, a mis en doute l'efficacité de l'analyse distributionnelle sur un corpus de langue générale du fait de sa grande variabilité syntaxique et lexicale. Dans un corpus de ce type, les restrictions de sélection entre un opérateur et son argument seraient trop floues, au contraire des corpus spécialisés (Harris, 1988).

Peu de travaux sur cette nouvelle orientation ont vu le jour. En français on peut citer ceux de Fabre, Habert et Labbé (1997), qui ont ainsi comparé les résultats de l'outil d'analyse distributionnelle automatique Zellig (Habert & Nazarenko, 1996) sur deux corpus : un corpus médical, constitué de comptes rendus d'hospitalisation, et un corpus plus général, constitué des discours politiques de François Mitterrand lors de son premier septennat. Il est apparu que si, pour le corpus spécialisé, les classes sémantiques induites sur la base de contextes syntaxiques partagés étaient bien délimitées, c'était moins le cas pour le corpus de discours politique. Les relations entre les mots étaient plus lâches, plus floues et très complexes du fait de la polysémie, de nombreux glissement de sens, etc.

Cette étude, loin de conclure sur l'inopérabilité de l'analyse distributionnelle en langue générale, indiquait les difficultés à prendre en compte lors de l'examen des ressources extraites sur des corpus plus généraux et montrait les limites de l'outil Zellig. L'une de celles-ci résidait dans le fait que l'analyseur syntaxique Lexter (Bourigault, 1994) utilisé en amont de Zellig se limitait à l'analyse des syntagmes nominaux. Or il semblait pour les auteurs que les verbes pourraient être un meilleur observatoire des proximités pour un corpus de langue générale.

Dans cet article, nous donnons les premiers résultats d'une étude exploratoire sur l'analyse distributionnelle de corpus de langue générale en nous intéressant plus

spécifiquement à la synonymie. Nous utilisons l'outil d'analyse distributionnelle automatique Upery (Bourigault, 2002) qui s'appuie sur les résultats de l'analyseur syntaxique Syntex, dont la couverture est beaucoup plus large que celle de Lexter (section 2). Nous appliquons cet outil sur deux corpus de grande taille : un corpus journalistique de 200 millions de mots et un corpus littéraire de 30 millions de mots (section 3). Nous procédons à une première évaluation des résultats en confrontant les voisins distributionnels aux synonymes du Dictionnaire Electronique de Synonymes de l'Université de Caen (section 4).

2. METHODE D'ANALYSE DISTRIBUTIONNELLE

L'outil d'analyse distributionnelle Upery s'appuie sur les résultats de l'analyseur syntaxique de corpus Syntex (Bourigault *et al*, 2005). Syntex prend en entrée un corpus de phrases étiquetées, et calcule pour chaque phrase les relations de dépendance syntaxique entre les mots (sujet, complément d'objet, complément prépositionnel, épithète, etc.). À partir de l'analyse syntaxique, Upery extrait des triplets (recteur, relation, régi). Par exemple, de l'analyse syntaxique de la phrase *« il mange la souris »*, il extrait le triplet *(manger, obj, souris)*. Au cours de cette étape d'extraction de triplets, sont effectuées automatiquement un certain nombre de normalisations syntaxiques, visant soit à passer de relations syntaxiques de surface à des relations syntaxiques plus profondes, soit à expliciter certaines relations (pour augmenter le nombre de relations détectées) :

- distribution de la coordination : on duplique les relations syntaxiques portant sur la coordination pour chacun des éléments coordonnés : *Jean mange la pomme et la poire* → *(manger, obj, pomme)* ; *(manger, obj, poire)* ;
- traitement du passif pour récupérer les objets profonds : *La pomme a été mangée* → *(manger, obj, pomme)* ;
- traitement de l'antécédence relative : *Jean qui dort* → *(dormir, suj, Jean)* ;
- recherche du sujet profond des verbes à l'infinitif (précédés de verbe à contrôle à l'actif) : si un verbe à l'infinitif est précédé d'un verbe à contrôle sujet, le sujet du verbe à l'infinitif sera le sujet du verbe à contrôle (ex : *Jean décide de nager* → *(nager, suj, Jean)*) ; si c'est un verbe à contrôle objet, le sujet de l'infinitif sera le complément du verbe à contrôle (ex : *Jean ordonne à Marie de travailler* → *(travailler, suj, Marie)*). Nous exploitons différentes listes de verbes à contrôle sujet ou objet qui se distribuent dans 8 patrons syntaxiques différents.

La méthode d'analyse distributionnelle que nous développons distingue deux types d'unités : les prédicats et les arguments. À chaque triplet extrait *(recteur, relation, régi)* correspond un couple *(prédicat, argument)* : le prédicat est constitué du recteur auquel on « accole » la relation, et l'argument est le régi. Les principales relations sont : *suj*, *obj*, les prépositions (*de*, *à*, *dans*, etc.), et *mod*, pour la relation épithète. Ainsi à partir du triplet *(mange, obj, souris)* on construit le couple *(manger_obj, souris)*. On procède ensuite à une double analyse distributionnelle : on rapproche les prédicats qui se construisent avec les mêmes arguments et on rapproche les arguments qui se construisent avec les mêmes prédicats. Par exemple, les prédicats *sculpter_obj* et *modeler_obj* sont rapprochés parce qu'ils se construisent de façon régulière avec les arguments *visage, corps, espace, personnage, forme, image*. De la même façon les arguments *tempête* et *tornade* sont

rapprochés parce qu'ils se construisent avec les prédicats *emporter_suj*, *déclencher_obj*, *résister_à*, *effet_de*, *provoquer_obj*. Nous disons que *(sculpter_obj, modeler_obj)* est un couple de « voisins prédicats » et que *(tempête, tornade)* est un couple de « voisins arguments ».

Pour calculer la proximité entre prédicats (ou entre arguments), nous faisons appel à la mesure du Jaccard qui fait intervenir le nombre de contextes partagés par les deux prédicats. Upery va calculer cette mesure de proximité sémantique pour tous les couples de prédicats et d'arguments qui partagent des contextes en commun. Ceux qui passent un certain seuil de proximité constitueront un couple de voisins.

3. DEUX CORPUS : LM10 ET FRANTEXT20

À l'heure actuelle, on ne dispose pas pour la langue française d'un corpus de référence du type du *Bristish National Corpus*. Pour initier cette étude, nous avons choisi de travailler sur deux corpus de grande taille, le critère du volume étant requis pour une meilleure efficacité des méthodes d'analyse distributionnelle (Grefenstette, 1994) :

- le corpus LM10 [1]**:** ce corpus est le recueil de 10 ans d'articles du Monde de 1991-2000. Il contient 200 millions de mots, ce qui représente à peu près 100 000 articles du journal. C'est un corpus journalistique abordant des thèmes variés ;
- le corpus Frantex20 [2]: il rassemble 515 romans du 20e siècle de la base Frantext, ce qui constitue environ 30 millions de mots.

Aucun des deux corpus n'est « représentatif » à lui seul de ce qu'est la langue générale : les mélanger ne conduirait pas plus à un corpus de référence. Avant de s'interroger sur la constitution d'un mega-corpus de référence en français, nous tenons à garder leur homogénéité pour pouvoir ensuite les contraster et faire émerger des usages différents au travers des résultats de l'analyse distributionnelle automatique, selon le point de vue d'analyse adopté (dans cet article : la synonymie).

Upery, l'outil d'analyse distributionnelle, a donc été appliqué sur ces deux corpus, séparément, ce qui a permis l'extraction de deux ressources lexico-distributionnelles : *Les voisins de LM10* et *Les voisins de Frantext20*. La taille de chacune des deux bases est décrite dans le tableau 1.

nombre	***Les voisins de LM10***	***Les voisins de Frantext20***
lemmes	145 164	45 428
couples de triplets	2 987 284	520 984
couples de voisins	6 183 862	792 354

Tableau 1 : Taille des deux bases

4. COMPARAISON AVEC LE DICTIONNAIRE ELECTRONIQUE DE SYNONYMES (DES)

4.1. Le DES

Nous présentons les premiers résultats d'une étude exploratoire sur les deux bases *Les voisins de LM10* et *Les voisins de Frantext20*. L'objectif est d'analyser de façon contrastive ces deux bases du point de vue de la synonymie. Intuitivement on peut s'attendre à ce que les rapprochements distributionnels rendent compte de relations synonymiques puisqu'ils font émerger des proximités sémantiques. De plus, les travaux sur l'acquisition automatique de relations sémantiques en corpus spécialisés offrent un état de l'art intéressant pour la détection de la synonymie, sur lequel on pourra s'appuyer (Hamon, 2000). Nous avons donc décidé de nous focaliser sur cette relation dans cette première comparaison des ressources extraites par Upery, en utilisant comme référence le *Dictionnaire Electronique de Synonymes* de Caen.

Le *Dictionnaire Electronique de Synonymes*[4], dont Jean-Luc Manguin est le responsable, a été développé au sein du laboratoire CRISCO. Le DES est la compilation de 7 dictionnaires classiques : Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse et Robert. Il contient 397 098 couples de synonymes. Chacune des relations synonymiques a été symétrisée, en d'autres termes si *A* a pour synonyme le mot *B* on ajoute le mot *A* dans les synonymes du mot *B*.

Dans cette première étude nous chiffrons le taux de recouvrement du DES avec chacune des deux bases et commentons de façon qualitative la mise au jour de certains cas de recouvrement et de non-recouvrement.

4.2. Un recouvrement faible : le DES / Les voisins

La difficulté du recouvrement, tant au niveau quantitatif que qualitatif, réside dans le fait que nous comparons des données de nature différente. Le DES contient des couples de mots, tandis que les voisins contiennent à la fois des couples d'arguments (qui ont la même forme que des mots) et des couples de prédicats (mot_ relation). Ainsi s'il y a recouvrement, un couple de mots dans le DES peut se trouver de une à plusieurs fois dans les voisins. Il peut ainsi correspondre soit à (cf. tableau 2) :

- un seul couple d'arguments comme le couple *[absurde, aberrant]* ;
- un seul couple de prédicats (cf. *[calmer, apaiser]*) ;
- un couple d'arguments et un couple de prédicats (cf. *[tempête, tornade]*) ;
- un couple d'arguments et à plusieurs couples de prédicats (cf. *[témoignage, déposition]*).

couples de synonymes du DES	**Voisins (LM10)** (**A**)rguments ou (**P**)rédicats		**Contextes**
[absurde, aberrant]	***A***	*(absurde, aberrant)*	*{logique_mod, comportement_mod, situation_mod, etc. }*
[calmer, apaiser]	***P***	*(calmer_obj, apaiser_obj)*	*{courroux, grogne, appréhension,, colère, etc}*
[tempête, tornade]	***A***	*(tempête, tornade)*	*{emporter_suj ; déclencher_obj, résister_à ; effet_de ; provoquer_obj }*
	P	*(tempête_mod, tornade_mod)*	*{médiatique violent, véritable, judiciaire, financier, nouveau }*
[témoignage, déposition]	***A***	*(témoignage, déposition)*	*{foi_de, recueillir_obj, conforter_suj, lecture_de, etc.}*
	P	*(témoignage_devant, déposition_devant)*	*{jury, commission, grand jury, etc.}*
		(témoignage_de, déposition_de)	*{témoin, expert, clinton, policier,etc.}*
		(témoignage_mod, déposition_mod)	*{accablant, nouveau, écrit, premier, etc.}*

Tableau 2 : Des synonymes aux voisins

Pour calculer le taux de recouvrement, nous n'avons pas tenu compte de ces doublons possibles dans les voisins : ainsi, quel que soit le nombre de fois où un couple apparaît dans les voisins il n'est compté qu'une fois.

Pour le corpus LM10, 22% des synonymes du DES se retrouvent dans *Les voisins de LM10*. Cela ne représente que 1% de voisins : c'est-à-dire que seul 1% de voisins sont des synonymes. Pour le corpus Frantext20, 10 % des synonymes du DES sont contenus dans *Les voisins de Frantext*, ce qui représente 3 % de voisins synonymes. Ce faible taux de recouvrement est une surprise. Certes en chiffre absolu, étant donné la taille des corpus, le nombre de relations extraites n'est pas négligeable : on extrait 50 848 couples de voisins synonymes pour LM10 et 18 704 pour Frantext20. Mais en terme de proportion, il y a très peu de relations synonymiques extraites par l'analyse distributionnelle par rapport à ce qu'on pouvait s'attendre au vu des travaux sur les corpus spécialisés, pour qui la relation synonymique était centrale - pas forcément selon des méthodes d'analyse distributionnelle

(Hamon & Nazarenko, 2001). Nous pouvons donc déjà constater un écart remarquable entre corpus spécialisé et corpus général.

Au-delà de ces chiffres, nous présentons trois cas faisant émerger certaines caractéristiques propres aux voisins par rapport à la synonymie. Les deux premières analyses montrent comment se manifeste la synonymie dans des résultats d'analyse distributionnelle automatique sur des corpus non spécialisés et les avantages qui en découlent. Le dernier cas s'arrête sur une limite de l'outil et les précautions qui s'ensuivent pour l'analyse.

4.3. Contextualisation de la synonymie

Cette première série de cas rend compte de l'apport des voisins dans la compréhension de la synonymie. Avec les voisins, une fois les synonymes connus, on accède à la contextualisation de la synonymie opérant dans le couple du DES. Cette mise en contexte se matérialise d'une part grâce aux notions de prédicats et d'arguments qui inscrivent les couples de voisins dans des relations syntaxiques particulières et d'autre part parce que, pour un couple donné, on a accès immédiatement aux régis ou aux recteurs des éléments du couple. Au travers de deux exemples de couples (cf. tableau 3) nous montrons l'intérêt d'une telle contextualisation.

couple de synonymes (DES)	corpus	couple de voisins (Les voisins)	contextes partagés (a=nombre)
[éprouver, ressentir]	LM	*(éprouver_obj, ressentir_obj)*	{*certain malaise, gêne, dégoût, honte, tristesse, besoin, compassion, soulagement, fascination, culpabilité, désir,* etc.} (a=61)
		(éprouver_suj, ressentir_suj)	{*lecteur, spectateur, visiteur, français, allemands, gens, américain, génération,* etc.} (a=34)
[souligner, insister]	LM	*(souligner_dans, insister_dans)*	{*entourage, discours, rapport, mode*} (a=4)
		(souligner_obj, insister_sur)	{*nécessité, importance, fragilité, importance de rôle,* etc.} (a=184)

Tableau 3 : Contextualisation de la synonymie

Comme on peut le voir dans le tableau 3, le couple de synonymes *[éprouver, ressentir]* s'actualise de deux façons dans les voisins. Les deux mots sont rapprochés par les objets

(éprouver_ obj, ressentir_obj) et par les sujets *(éprouver_suj, ressentor_suj)* qu'ils partagent. Les mots du couple ne sont plus des unités lexicales *in abstracto* comme dans le DES, on voit dans quel contexte ces deux mots sont synonymes. De cette façon, les voisins nous éclairent sur la notion de synonymie partielle : *éprouver* et *ressentir* ne sont pas synonymes de manière absolue mais seulement quand ils sont engagés dans les relations sujet ou objet - ces relations ont certes une large couverture, mais ce n'est pas le cas pour tous les recouvrements (*cf.* le couple *[souligner, insister]*).

L'intérêt de la contextualisation est plus probant encore avec le couple *[souligner, insister]*, synonymes dans le DES, rapprochés eux aussi de deux façons différentes dans les voisins et dont les deux mots s'inscrivent dans des relations syntaxiques moins classiques. En premier lieu, ce sont les compléments prépositionnels en *dans* qui permettent de faire émerger le couple. Il peut paraître surprenant de rapprocher des verbes sur la base de compléments circonstanciels partagés, compléments qui n'entrent pas traditionnellement dans le cadre de sous-catégorisation des verbes. Or les faits sont là et l'examen d'autres couples nous montre que ce cas est loin d'être unique. Cette possibilité de rapprochement est offerte par Syntex qui ne fait pas de distinction entre les compléments essentiels et non essentiels. Il permet donc de récupérer des couples de voisins rapprochés pour des raisons qui semblent de prime abord moins pertinentes mais qui sont loin d'être anecdotiques et qu'il faudra creuser plus avant dans la suite de nos études.

Le deuxième couple de prédicats faisant intervenir *[souligner, insister]* est également intéressant. En effet comme on peut le voir dans le tableau, les deux prédicats ne comportent pas la même relation. Pour l'un (*souligner*) les contextes partagés sont en position objet (*souligner_obj*) alors que pour l'autre (*insister*) les mêmes contextes sont précédés de la préposition *sur* (*insister_sur*). Pour autant, même si les relations ne sont pas identiques, le couple *(souligner_obj, insister_sur)* est bien un couple de synonymes (les contextes partagés nous le prouvent). Les voisins font ici apparaître une dissymétrie que les synonymes dans le DES ne laissent pas supposer. En effet, on aurait tendance à croire que deux mots synonymes sont substituables dans le même contexte or ici le contexte droit est tout à fait différent. La contextualisation de la synonymie prend donc toute son importance. On peut raisonnablement imaginer que le lexicographe qui a statué sur la synonymie de *souligner* et *insister* a envisagé cette contextualisation : *souligner_obj* est synonyme de *insister_sur* et on n'admettrait pas une symétrie dans la relation de type (*souligner_obj, insister_obj*) : on n'insiste pas X, on insiste sur X. Le format du DES ne nous permet pas de voir ce cadre de sous-catégorisation, il présente les synonymes hors de tout contexte et induit même une symétrie, parfois abusive. Ce sont donc les voisins qui permettent d'accéder, une fois les synonymes connus, à la façon dont les mots sont synonymes en contexte (et relativement à ce corpus).

L'usage des voisins, peut ainsi s'avérer complémentaire sinon nécessaire à une recherche dans le DES puisqu'ils donnent accès à des informations sur les restrictions de sélection de chacun des termes du couples que les dictionnaires de manière générale ne fournissent pas. Ces propriétés syntaxiques conditionnent le diagnostic de synonymie comme en témoignent les deux exemples analysés ci-dessus.

4.4. La synonymie sensible aux usages

Ce second cas de recouvrement (et de non-recouvrement) montre l'intérêt de travailler sur des ressources différentes : la confrontation des deux bases *Les voisins de LM10* et *Les voisins de Frantext20* nous montre combien la synonymie est sensible aux genres et aux thèmes abordés. En effet selon la base utilisée on retrouve ou non certains synonymes du DES.

Le cas du couple *[abattement, désespoir]* est pour cela exemplaire. Il s'agit d'un couple de synonymes du DES qui n'apparaît pas dans *Les voisins de LM10* mais qu'on retrouve dans *Les voisins de Frantext20* sous la forme d'un couple d'arguments (*abattement, désespoir).* L'absence de ce couple dans *Les voisins de LM10* semble s'expliquer en confrontant ce couple à un autre : *[abattement, exonération]*, couple de synonymes dans le DES et présent sous diverses formes dans *Les voisins de LM10* mais absent, cette fois, dans *Les voisins de Frantext20* (cf. tableau 4).

couple de synonymes (DES)	corpus	couple de voisins (Les voisins)	contextes partagés (a=nombre)
[abattement, désespoir]	LM	∅	∅
	FR	*(abattement , désespoir)*	*{état_de, pouvoir_suj, forme_de, moment_de}* (a=4)
[abattement, exoneration]	LM	*(abattement, exoneration)*	*{bénéficier_de, plafond_de, consentir_obj, limiter_obj, montant_de, taux_de, système_de, etc.}* (a= 30)
		(abattement_de, exonération_ de)	*{cotisation patronale, charge sociale, cotisation sociale, taxe professionnelle, impôt, etc.}* (a=8)
		(abattement_mod, exonération_mod)	*{fiscal, supplémentaire, important, etc.}* (a=6)
	FR	∅	∅

Tableau 4 : Différences d'usage selon les corpus

Avec ces deux couples de synonymes du DES, deux sens de *abattement* sont instanciés. En effet, selon *Le Robert*, *abattement* a deux sens : soit il s'apparente à une réduction et dans ce cas *abattement* est synonyme d'*exonération [abattement, exonération]*, soit il désigne l'état d'une personne (psychologique ou physique) et il est alors synonyme de *désespoir [abattement, désespoir]*.

Dans le corpus LM10 c'est le sens de *abattement* comme « réduction », qui prédomine. En témoigne la présence du couple *[abattement, exonération]* que l'on retrouve sous diverses formes dans les voisins de LM : (*abattement, exonération*) (*abattement_de, exonération_de)* et (*abattement_mod, exonération_mod).* Les contextes de chacun de ces couples de voisins nous permettent de confirmer l'actualisation du premier sens de *abattement*. De plus, les autres couples de synonymes du DES qui mettent en jeu *abattement* et que l'on retrouve dans *Les voisins de LM10* sélectionnent tous ce sens là : [*abattement, dégrèvement], [abattement, allègement], [abattement, réduction]*. On ne trouve pas de synonyme sélectionnant un autre sens de *abattement* dans cette base. Le couple *[abattement, dépression]* n'est donc pas présent dans les voisins de LM10. Ceci n'est pas étonnant vu la nature journalistique du corpus LM10 qui traite sûrement plus souvent de sujets économiques ou boursiers que de l'état psychologique des individus.

Dans Frantext20 par contre, c'est le sens de *abattement* comme état dépressif ou physique qui est représenté : le couple de synonymes du DES *[abattement, dépression]* est cette fois présent dans la base, sous la forme de *(abattement, désespoir)*. L'autre couple du DES, comprenant le terme *abaissement* et présent dans *Les voisins de Frantext20*, fait d'ailleurs lui-aussi clairement référence à l'état psychologique : *[abaissement, dépression]*. Comme attendu, le premier sens d'*abattement* est absent : on ne trouve pas le couple [*abattement, exonération]* ce qui n'est pas surprenant, Frantext20 étant un corpus littéraire.

Après avoir considéré l'intérêt d'inclure une dimension syntaxique et contextuelle dans la description de synonymie (cf. section 4.2), nous saisissons donc, au travers de cette série de recouvrement, la pertinence d'une prise en compte de la diversité des discours : à la différence des ressources dictionnairiques, les voisins permettent, en naviguant d'un corpus à l'autre, de préciser des informations sur le registre discursif ou thématique qui viennent enrichir la description des synonymes.

4.5. Synonymie partielle des contextes

Dans ce dernier cas de recouvrement, nous présentons un problème lié à l'utilisation d'un outil d'extraction automatique et proposons une précaution méthodologique pour l'analyse : certains couples de voisins, correspondant à des synonymes, ne sont pas synonymes dans tous les contextes qui ont permis leur rapprochement, il est donc nécessaire de vérifier la pertinence des rapprochements grâce à la visualisation des contextes. Nous illustrons cette limite avec le cas de *[embrasser,voir]*.

Soit le couple de synonymes du DES *[embrasser, voir]*. En se plaçant du point de vue de *embrasser*, l'examen de ses différents sens explique la relation synonymique qu'il entretient avec *voir*. En effet, Dans *Le Robert*, *embrasser* a 4 sens :

1) prendre et serrer dans ses bras
2) donner un baiser à quelqu'un
3) adopter, embrasser la carrière de
4) saisir par la vue dans toute son étendue – par analogie (comprendre, concevoir)

C’est dans ce dernier sens, qu’*embrasser* est synonyme de *voir*.

couple de synonymes (DES)	corpus	couple de voisins (Les voisins)	contextes partagés (a=nombre)
[embrasser, voir]	LM	*(embrasser_obj, voir_obj)*	{*petite fille, bébé, dame, totalité, horizon, champ, carrière, cause, garçon, paysage, univers, monde*, etc.} (a=40)

Tableau 5 : Synonymie partielle des contextes

Comme on peut le voir dans le tableau 5, le couple *[embrasser, voir]* est bien présent dans la base *Les voisins de LM10*. Il se manifeste sous la forme de deux voisins prédicats : (*embrasser_obj; voir_obj*). Toutefois les contextes partagés qui ont permis le rapprochement de ces deux éléments ne sont pas tous pertinents au regard de cette relation synonymique. En effet, pour *embrasser*, certains contextes sélectionnent le sens 1 ou le sens 2 (prendre et serrer dans les bras ou donner un baiser à quelqu’un). C’est le cas de *{petite fille, bébé, dame, garçon}*. Une autre série de contextes actualise le sens 3 : *{carrière, cause}*. Finalement, seuls 6 contextes sur 40 semblent bien sélectionner le sens 4 pour lequel *embrasser* est synonyme de *voir* : *{totalité, horizon champ, paysage, univers, monde}*.

Ce premier constat nous incite donc à être vigilants par rapport aux contextes permettant le rapprochement de deux voisins : même si le couple de voisins extrait semble pertinent, les raisons sous-jacentes à ce rapprochement sont parfois partiellement fortuites. Partiellement, car nous n’avons pas trouvé de cas où le rapprochement était une pure coïncidence : il y a toujours au moins quelques contextes qui actualisent le sens pour lesquels les deux termes sont synonymes.

Pour *(embrasser obj, voir obj)*, plusieurs explications à leur rapprochement dans le cas des contextes non pertinents sont possibles. On peut imaginer que si ces prédicats ont été rapprochés c’est parce que les objets de *embrasser* et *voir* partagent au moins un sème en commun ou plus simplement que ce qu’on peut *embrasser* on peut également le *voir*. Soit dans les sens 1 et 2 parce que le contexte fait référence à un objet ou un sujet physique, matérialisé (on peut voir une fille et on peut l’embrasser). Soit dans le sens 3 parce que le contexte renvoie à un concept et cette fois ce qu’on peut *embrasser* dans le sens de « adopter » on peut aussi le *voir*, « l’imaginer » (*carrière, cause*).

Avec ce cas qui n’est pas isolé, on s’aperçoit donc que tous les contextes permettant de rapprocher deux unités ne sélectionnent pas forcément le même sens : il n’y a pas obligatoirement d’homogénéité sémantique. Ceci s’explique par la polysémie des unités lexicales qui n’est pas contrôlée dans Upery et qui est très fréquente en langue générale. On retrouve donc ici, à l’intérieur même d’un couple, un problème soulevé par Habert, Fabre et Labbé (97). Il est de ce fait indispensable pour l’utilisateur des bases de vérifier pourquoi un couple a été rapproché – et donc d’opérer un véritable retour au contexte – afin d’en saisir la pertinence.

5. CONCLUSION

Dans cet article nous avons montré des résultats de l'analyse distributionnelle automatique, au moyen de l'outil Upery, appliqué à des corpus non spécialisés : LM10 et Frantext20. Ces premiers résultats se présentent sous la forme de deux bases lexico-distributionnelles : *Les voisins de LM10* et *Les voisins de Frantext20*. Après avoir expliqué la problématique qui sous-tend le choix de corpus généraux et la méthode mise en œuvre pour faire de l'analyse distributionnelle, nous avons présenté une première étude exploratoire sur la synonymie et comparé à ce titre *Les voisins de LM10* et *Les voisins de Frantext20* avec le *Dictionnaire Electronique de Synonymes* de Caen.

La projection des couples de synonymes du DES sur les couples de voisins nous a permis au travers de quelques cas de recouvrement, de faire émerger certaines caractéristiques propres aux voisins face à d'autres ressources dictionnairiques :

- la possibilité avec les voisins d'accéder à la contextualisation des couples de synonymes présentés *in abstracto* dans le DES ;
- la synonymie étant sensible aux usages, le fait de disposer de descriptions différentes des relations synonymiques selon les thématiques abordées (et donc selon les corpus) ;
- la nécessité de rester prudent face aux rapprochements extraits des voisins, dans la mesure où les contextes qui permettent le rapprochement de couples de synonymes ne sont parfois que partiellement pertinents.

Nous insistons sur le fait que ces quelques points ne sont que les résultats d'une première exploration des voisins et qu'ils méritent d'être creusés plus longuement. Si la synonymie semble bien se manifester aux travers des proximités distributionnelles, il faudrait maintenant organiser une typologie de ses différentes manifestations, décrire ainsi des comportements plus systématiques et relever le cas échéant des indices formels permettant la détection automatique des liens synonymiques.

Par ailleurs les synonymes, comme on l'a vu, ne représentent que 1 % à 3 % des voisins. Qu'en est-il des autres ? Une piste serait de recenser et d'étudier les différentes relations sémantiques que les proximités distributionnelles mettent au jour comme l'hypéronymie, l'antonymie, la méronymie, etc. et, à terme, de mettre en œuvre des moyens de les détecter automatiquement.

6. NOTES DE FIN

1. Le corpus LM10 a été construit à partir de ressources obtenues auprès de l'agence ELDA, à l'aide de programmes de nettoyage et de balisage réalisés par B. Habert du LIMSI, que nous remercions ici.
2. Fourni par l'ATILF, grâce à une convention ATILF / ERSS
3. http://www.ims.uni-stuttgart.de
4. http://elsap1.unicaen.fr/dicosyn.html

7. RÉFÉRENCES

Bourigault D. (1994) *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*, Thèse, École des Hautes Études en Sciences Sociales, Paris.

Bourigault D. (2002) 'Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologie à partir de corpus', dans *Actes TALN' 2002*, Nancy.

Bourigault D., Aussenac-Gilles N., Charlet J. (2004) 'Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas', *Revue d'Intelligence Artificielle (RIA)*, 'Techniques Informatiques et structuration de terminologiques', Pierrel J.-M. et Slodzian M. (Ed.), Paris : Hermès. Vol. 18, n°1/2004, p 87-110.

Bourigault D., Fabre C., Frérot C., Jacques M.-P. & Ozdowska S. (2005), 'Syntex, analyseur syntaxique de corpus', dans *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan.

Grefenstette G. (1994) *Explorations in Automatic Thesaurus Discovery*, Londres, Kluwer Academic Publishers.

Habert B., Fabre C., Labbé D. (1997) 'La polysémie dans la langue générale et les discours spécialisés', *Sémiotiques*, n°13, 1997, p. 15-31.

Habert B., Nazarenko A. (1996) 'La syntaxe comme marche-pied de l'acquisition de connaissances : bilan critique d'une expérience', dans *Actes des Journées d'acquisition des connaissances (JAC 96)*, p. 137-149.

Hamon T. (2000) *Variation sémantique en corpus spécialisés : Acquisition de relations de synonymie à partir de ressources lexicales*. Thèse, Université Paris-Nord, Villetaneuse.

Hamon T., Adeline Nazarenko (2001) 'Detection of synonymy links between terms: experiment and results', dans Bourigault, D., Jacquemin, C., L'Homme, M.C. (eds) *Recent Advances in Computational Terminology*. John Benjamins.

Harris Z. (1968) *Mathematical Structures of Language*, John Wiley & Sons, New York.

Harris Z. (1988) *Language and information*. Columbia University Press, New York:

Sager N., **Friedman C., Lyman M. S.** (1987) *Medical Language Processing : Computer Management of Narrative Data*, Addison-Wesley.

ÉTUDE SIMULTANÉE DE LA SYNONYMIE ET DE L'ANTONYMIE EN CONTEXTE PAR ANALYSE DISTRIBUTIONNELLE CONTRAINTE

Jean-Luc MANGUIN (1) ; Itsuko FUJIMURA (2) ;
Guillaume JACQUET (3) ; Fabienne VENANT (3)
(1) CRISCO, CNRS & Université de Caen ;
(2) Université de Nagoya ; (3) LATTICE, CNRS & Université de Paris 61.

OBJECTIFS DE L'ÉTUDE

Cette étude se propose d'analyser et de représenter le sens d'un adjectif du lexique par un objet géométrique visualisable, dont les différentes régions correspondent aux différents sens pris en contexte nominal. Nous avons déjà montré (Manguin, à paraître) qu'il est possible d'établir une cartographie sémantique d'une telle unité lexicale à partir de ses propres cooccurrences nominales et de celles de ses synonymes, recueillies dans un corpus, pourvu que l'on applique à celles-ci certaines restrictions décrites dans le paragraphe suivant ; la représentation obtenue permettait d'observer le déploiement du sens de l'adjectif étudié, et de constater quels étaient les contextes associés à ces différentes acceptions. Cette fois, nous voulons répondre à la question de la stabilité de cette représentation en fonction du corpus d'observation, en particulier nous désirons savoir si les contextes nominaux (ou peut-être les « classes » de contextes) sont répartis de la même manière. Et d'autre part, nous souhaitons diminuer les contraintes de cette analyse distributionnelle particulière, en incluant les antonymes de l'adjectif étudié dans l'ensemble paradigmatique employé lors de la première partie de cette étude.

2. MÉTHODE ET CORPORA

2.1. Méthode

Notre méthode d'analyse des cooccurrences est un processus en plusieurs étapes, dans lequel les résultats obtenus à un certain stade doivent être transmis au traitement suivant. Comme nous l'avons dit, une fois que nous avons choisi l'adjectif à étudier, nous recherchons dans le corpus quels sont les noms dont il peut être épithète, autrement dit nous devons repérer dans le corpus les occurrences « NOM + ADJ » ou « ADJ (+ ADV) + NOM », où ADJ est notre adjectif dans toutes ses formes possibles ; cette manière de procéder implique que le corpus soit au moins balisé, avec indication de la catégorie et du lemme ; le balisage fonctionnel n'est pas nécessaire, mais d'autre part les erreurs éventuelles de l'analyseur qui pose ces balises nous obligent à vérifier les exemples recueillis pour les débarrasser des fautes de catégorisation automatique. À l'issue de cette étape, nous possédons la liste des noms qui, dans le corpus étudié, cooccurrent avec notre adjectif ; nous ne gardons de cette liste que les éléments les plus fréquents, soit normalement ceux dont les exemples représentent plus de 0,5 % du total des cooccurrences. Nous voyons déjà que nous avons imposé dès cette première étape une contrainte à l'analyse distributionnelle que nous effectuons : cette contrainte est

catégorielle, puisque nous imposons au cooccurrants de notre unité d'appartenir à la catégorie nominale.

L'étape suivante voit l'apparition de la seconde contrainte, cette fois d'ordre paradigmatique : nous allons à présent relever les cooccurrences de même structure que précédemment, mais en exigeant que NOM soit un membre de la liste constituée à la première étape, et que ADJ soit un synonyme ou un antonyme de l'adjectif étudié ; c'est dans cette deuxième condition que réside la contrainte supplémentaire de notre analyse distributionnelle. Le choix des synonymes et des antonymes se fait au moyen d'un dictionnaire élaboré au laboratoire Crisco, et déjà décrit dans de précédentes publications. Le principe sous-jacent à cette contrainte est que la substituabilité de l'unité étudiée par un de ses synonymes est liée au contexte nominal ; en outre, le fait d'inclure les antonymes permet d'élargir la substitution synonymique au remplacement par un antonyme précédé d'un adverbe de négation ou d'affaiblissement.

Le traitement automatique des exemples recueillis permet de disposer de paires adjectivo-nominales et de leurs fréquences absolues dans le corpus, que nous pouvons présenter sous forme d'un tableau NOM vs ADJ, dans lequel le contenu de chaque case sera la fréquence de la paire correspondante ; nous transformons alors l'ensemble du tableau en fréquences relatives, en divisant chaque fréquence absolue par la fréquence absolue (dans le corpus) de l'adjectif impliqué dans la paire. Ceci nous permet de nous affranchir des fréquences adjectivales, et de comparer les adjectifs uniquement par les proportions d'emploi avec les noms de la liste.

La dernière étape consiste à appliquer à ces données une « métrique du chi2 », puis à effectuer sur ces nouvelles valeurs une analyse factorielle des correspondances, ce qui nous donne les coordonnées des différents points de la représentation. Ce traitement final s'inspire directement du processus décrit dans Jacquet & Venant (2005).

2.2. Corpora

Comme nous venons de l'indiquer, nous utilisons des corpus balisés, assez larges afin de pouvoir recueillir un nombre suffisant d'exemples. Le premier corpus est la base Frantext catégorisée, qui s'étend de 1830 à nos jours, et dont nous avons exclu les textes poétiques ; il demeure 1711 textes, représentant un total de 120 millions de mots. Le second corpus est un ensemble de textes de journaux et de magazines contemporains : *Le Monde* (1987 – 1992), *Le Monde Diplomatique* (1984 – 1998), *Libération* (1995 – 1999) et *Le Point* (1995 – 2000) ; les textes comptent 208 millions de mots au total, et ils sont tous balisés par le logiciel TreeTagger. Ce corpus a déjà été utilisé dans d'autres publications (voir Fujimura, Uchida & Nakao, 2004).

3. RÉSULTATS

Rappelons que cette étude prolonge les travaux déjà présentées en 2003 lors des précédentes journées de linguistique de corpus de Lorient ; c'est la raison pour laquelle nous avons décidé de continuer sur le même adjectif, en l'occurrence *curieux*, qui possède la particularité déjà signalée (voir Bat-Zeev Schyldkrot, 1997 ; François, Victorri & Manguin, 2003) de changer de sens en changeant de position par rapport au substantif dont il est l'épithète ; ce phénomène se traduit assez clairement, comme nous le verrons, sur la forme de son espace sémantique.

3.1. Contextes nominaux

Nous donnons ci-dessous les contextes nominaux obtenus pour chacun des corpus, suivant la méthode exprimée plus haut ; nous avons indiqué en gras les contextes partagés par les deux ensembles de textes, afin de les conserver comme points de comparaison.

Les contextes issus de Frantext (25) :

> ***bête***, *cas*, ***chose***, *conversation*, *détail*, ***esprit***, *exemple*, *expérience*, *fait*, *figure*, ***histoire***, ***livre***, ***mélange***, ***objet***, ***oeil***, *page*, ***personnage***, ***phénomène***, ***regard***, ***spectacle***, *spécimen*, *travail*, *type*, *tête*, *étude*.

Les contextes issus des journaux et magazines (55) :

> *affaire*, *alchimie*, *alliance*, *ambiance*, *atmosphère*, *attitude*, ***bête***, *chassé-croisé*, ***chose***, *cocktail*, *comportement*, *conception*, *couple*, *coïncidence*, *effet*, ***esprit***, *exercice*, *façon*, *film*, *forme*, *gens*, *habitude*, ***histoire***, *homme*, *idée*, *impression*, *lecteur*, ***livre***, *manière*, ***mélange***, *méthode*, *nom*, ***objet***, ***oeil***, *opération*, *paradoxe*, *pays*, ***personnage***, ***phénomène***, *pratique*, *public*, *rapport*, ***regard***, *renversement*, *retour*, *retournement*, *roman*, *résultat*, *sensation*, *sentiment*, *situation*, ***spectacle***, *système*, *tour*, *visiteur*.

Nous remarquons dès à présent que de nombreux contextes trouvés par Frantext apparaissent dans la liste de ceux trouvés dans le corpus de journaux et magazines, et que parmi ces contextes communs, certains comme *regard* permettent à *curieux* de changer de sens suivant sa position (*un curieux regard* vs *un regard curieux*).

3.2. Représentation avec synonymes

Dans la suite du processus, nous recherchons dans nos corpus les cooccurrences des différents noms avec les synonymes de *curieux* pris dans notre dictionnaire des synonymes du CRISCO ; dans le cas présent, les synonymes retenus sont :

> *amusant*, *anxieux*, *attachant*, *attentif*, *avide*, *bizarre*, *chercheur*, *drolatique*, *drôle*, *déconcertant*, *extraordinaire*, *extravagant*, *fureteur*, *incompréhensible*, *inconcevable*, *incroyable*, *indiscret*, *inouï*, *inquisiteur*, *intéressant*, *intéressé*, *investigateur*, *original*, *paradoxal*, *piquant*, *pittoresque*, *plaisant*, *rare*, *singulier*, *soucieux*, *surprenant*, *unique*, *étonnant*, *étrange*.

Quel que soit le corpus utilisé, les exemples recueillis (plusieurs milliers, vérifiés manuellement) montrent que toutes les associations possibles n'apparaissent pas dans les textes ; par exemple, dans le cas du corpus Frantext, sur les 1750 paires possibles, on n'en observe que 414 qui sont réalisées.

Nous donnons aux figures 1 et 2 les représentations obtenues respectivement avec Frantext et avec le corpus de presse contemporain ; la ressemblance des deux figures est frappante : dans les deux cas, les synonymes se regroupent avec une grande pertinence sémantique, mais en outre, les adjectifs sont clairement dissociés en deux groupes : à droite le sens de curieux « intéressé » et à gauche celui de curieux « intéressant » ; qui plus est, la zone de gauche montre la même évolution géométrique du « surprenant » et de l'« inexplicable » (en haut) jusqu'au « pittoresque » et à l'« attachant » (en bas). Par souci de lisibilité, nous n'avons pas étiqueté la zone centrale, mais les figures complètes de notre article peuvent être consultées sur http://www.crisco.unicaen.fr/.

Au point de vue des contextes, les plus agentifs (*œil*, *regard*, voire *esprit*) se placent logiquement à droite ; ceux qui se placent à gauche montrent que leur emploi avec *curieux*

les place dans une classe de patients ; ainsi *personnage* n'est pas représenté parmi les agents. De plus, ces contextes « patients » sont associés aux synonymes de la même manière dans les deux corpus : dans la zone de l'« inexplicable », on trouve des contextes génériques comme *phénomène*, *chose*, *cas*, *situation* ou encore *effet*, tandis que les contextes désignant des entités précises se rencontrent dans la partie du « pittoresque », comme *personnage*, *histoire*, *livre*, *détail*, ou *roman*. L'explication de ces placements est relativement simple : ce qui est *inexplicable*, *surprenant* ou *paradoxal* est aussi difficilement reconnaissable (peut-être à cause de l'émotion du moment) et par conséquent difficile à nommer ; à l'opposé, ce qui est *plaisant* et *amusant* est facilement caractérisé et donc nommable.

Globalement, nous pouvons dire que dans le cas des cooccurrences avec synonymes seuls, il n'y a pas de différence de représentation en faisant varier le corpus, ce qui en soit constitue un résultat intéressant.

3.3. Représentation avec synonymes et antonymes

Nous avons répété notre analyse en relevant ensuite les occurrences des noms de la liste et d'un antonyme de *curieux* ; les antonymes retenus sont :

> *anodin, banal, commun, courant, discret, habituel, indifférent, insouciant, ordinaire, quelconque, réservé, usuel, vulgaire.*

Les représentations obtenues sont données par les figures 3 et 4, et l'on remarque une certaine déformation de la disposition, sans que l'on puisse parler de bouleversement (les antonymes sont repérés par l'encadrement de leur étiquette).

La visualisation obtenue avec le corpus de périodiques est la moins déformée, et le placement des antonymes est des plus logiques ; par rapport à la figure 2, on remarque que l'axe vertical est en quelque sorte « inversé », et que la plupart des antonymes se trouvent dans la zone de gauche qui correspond aux contextes « patients ». Seul *indifférent* se situe dans la zone des agents, ce qui correspond à son acception habituelle. Le placement détaillé des autres antonymes n'est pas toujours en correspondance avec leurs opposés parmi les synonymes ; on peut toutefois signaler que *réservé* et *courant* sont plutôt logiquement situés.

La représentation issue du corpus Frantext appelle plus de commentaires ; tout d'abord, les antonymes situés dans la zone de gauche, comme *anodin*, *quelconque*, *vulgaire*, *usuel*, *courant* ou *habituel* possèdent ici un emplacement plus proche de leur(s) opposé(s) parmi les synonymes ; par contre, *réservé* et *insouciant* sont assez excentriques, ce qui peut mettre en doute leur emploi d'antonyme de *curieux* dans leurs contextes préférentiels qui sont *conversation réservée* et *esprit insouciant*. D'autre part, la position de *discret* à mi-chemin des deux zones « patients » et « agents » peut laisser penser que son emploi dans *regard discret* est parfois antonymique de *regard indiscret*, mais l'examen des exemples recueillis infirme cette hypothèse ; nous touchons là les limites de notre méthode représentative. Enfin, la position de *indifférent* peut surprendre : situé lui aussi à mi-chemin des deux zones, il semble qu'il puisse qualifier des agents comme des patients ; cette fois, les exemples révèlent que notre représentation n'est pas trompeuse, puisque des exemples de *choses indifférentes* (c'est-à-dire de choses qui laissent les gens sans réaction) abondent dans la littérature du 19ème siècle, et reflètent un emploi soutenu qui n'est plus d'actualité.

Ainsi notre méthode permet de mettre en lumière certaines évolutions diachroniques lors des comparaisons entre corpus de différences époques ; cela dit, son « mécanisme » ne

doit pas être perdu de vue, sous peine de conduire à des interprétations fausses en ce qui concerne les liens paradigmatiques (comme dans le cas de *discret*).

4. BILAN ET PERSPECTIVES

À l'issue de cette étude, nous pouvons affirmer que la méthode que nous avons mise en œuvre s'avère donner des résultats stables malgré les changements de corpus ; de plus, la stabilité globale ne gomme pas les variations qui caractérisent les emplois de l'unité étudiée selon les textes retenus pour l'application du traitement. En revanche, l'élargissement paradigmatique opéré n'a pas tout à fait rempli nos espérances, et ceci pour deux raisons : la première vient du fait que l'hypothèse de substituabilité formulée au départ ne se vérifie pas dans la réalité, autrement dit on rencontre assez peu d'adjectifs épithètes précédés d'un adverbe de négation ou d'affaiblissement. Dans notre travail, nous n'avons pas cherché de tels cas de substitution, mais plutôt les emplois antonymiques, cependant il est probable que l'application de notre méthode au cas de l'adjectif attribut serait probablement plus fructueuse, mais les corpus employés nécessiteraient alors un balisage plus élaboré, qui devrait résoudre les problèmes de référence, vu le nombre d'occurrences où l'adjectif est attribut d'un pronom. Par ailleurs, il n'est pas exclu qu'une différence de nature entre la synonymie et l'antonymie conduise, lors de la transformation d'un énoncé en son contraire, à une double substitution du nom et de l'adjectif, par exemple *un homme intelligent* devient *un gars bête* (ou *un type bête*) ; ceci peut expliquer l'absence de certaines occurrences escomptées.

La deuxième raison réside dans la polysémie des adjectifs, qui est un problème déjà rencontré avec *sec* dans un traitement automatique proche du nôtre (voir Venant, 2004). Pour le résumer, disons simplement qu'un adjectif synonyme de *sec* pour un contexte donné peut se rencontrer avec un autre contexte sans qu'il y ait cette fois de synonymie ; par exemple : *accueil sec* / *accueil glacial*, comparé à *temps sec* / *temps glacial*. Dans notre cas, nous pouvons rencontrer *personnage commun*, mais aussi *histoire commune*, où cette fois l'antonymie avec *curieux* est loin d'être certaine.

Ainsi, les résultats prometteurs obtenus dans la direction de la comparaison de corpus au travers d'une unité lexicale devront être complétés, dans le domaine de l'étude antonymique, par une réflexion méthodologique et des outils plus élaborés.

5. RÉFÉRENCES

Bat-Zeev Schyldkrot H., 1997. 'Synonymie et polysémie, le cas de CURIEUX comme parcours sémantique d'un mot', *Langages*, 128, p. 113-125.

François J., Victorri B. & Manguin J.-L., 2003. 'Polysémie adjectivale et synonymie : l'éventail des sens de curieux', in Soutet O. (ed.) *La polysémie*, Paris : Presses de l'Université de la Sorbonne.

Frantext, base textuelle catégorisée, 1999. CNRS, ATILF (Analyse et traitement informatique de la langue française), UMR CNRS-Université Nancy2, http://www.inalf.fr/atilf .

Fujimura, I., Uchida, M., Nakao H., 2004. '*De* vs *des* devant les noms précédés d'épithète en français : le problème de *petit*', in *Le poids des mots, Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles*, Louvain La Neuve.

Jacquet G., Venant F., 2005 . 'Construction automatique de classes de sélections distributionnelles', *12ème conférence annuelle sur le traitement automatique des langues* (TALN 05), Dourdan.

Manguin J.L., (à paraître). 'Utilisation d'un corpus catégorisé pour l'étude et la représentation de la synonymie en contexte' in Williams G. (ed.), *Actes des 3èmes journées de linguistique de corpus*, Paris : L'Harmattan.

Venant F., 2004. 'Polysémie et calcul du sens' in *Le poids des mots, Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles*, Louvain La Neuve.

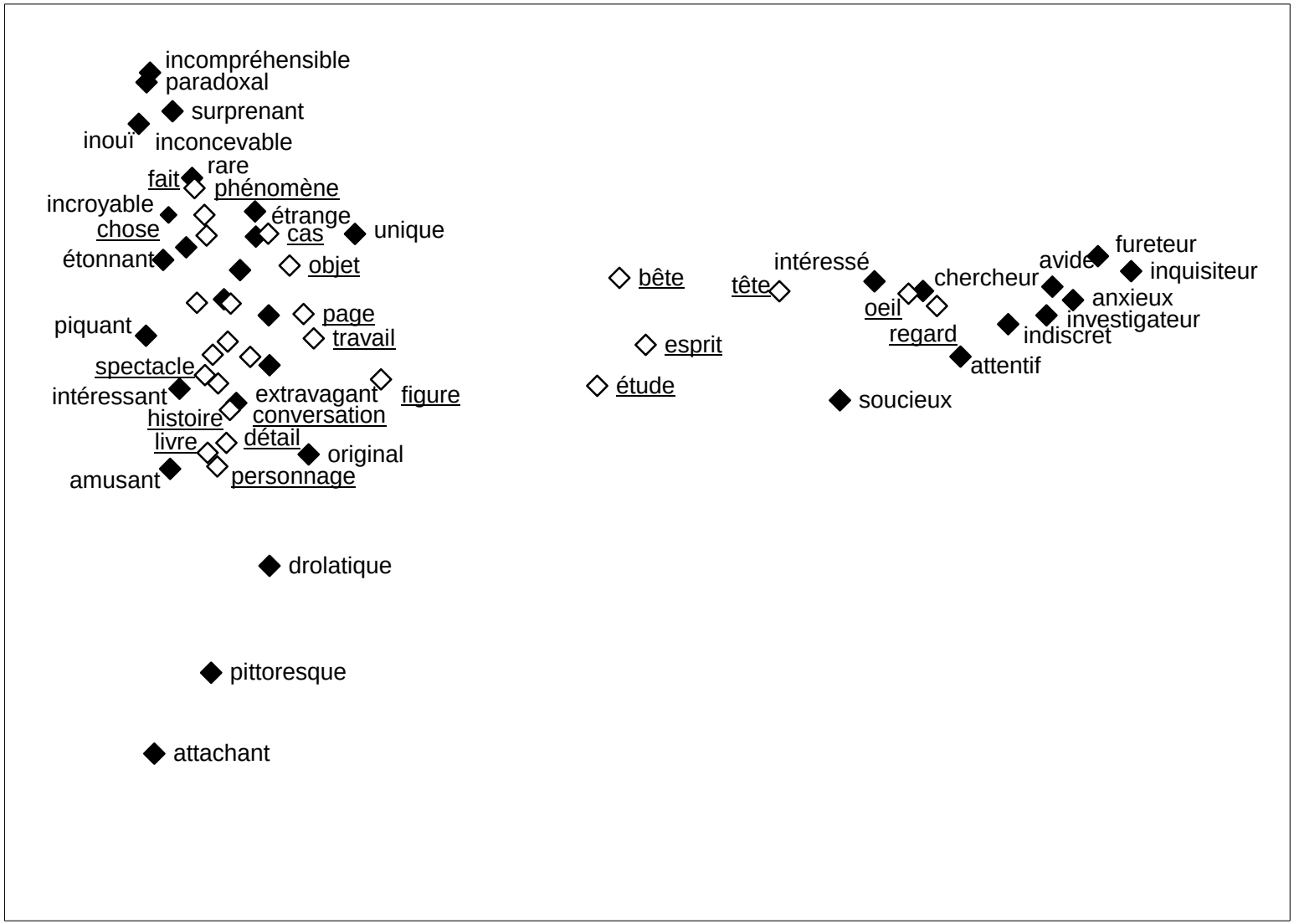


Figure 1 : espace sémantique obtenu avec les synonymes et les cooccurrences dans le corpus Frantext

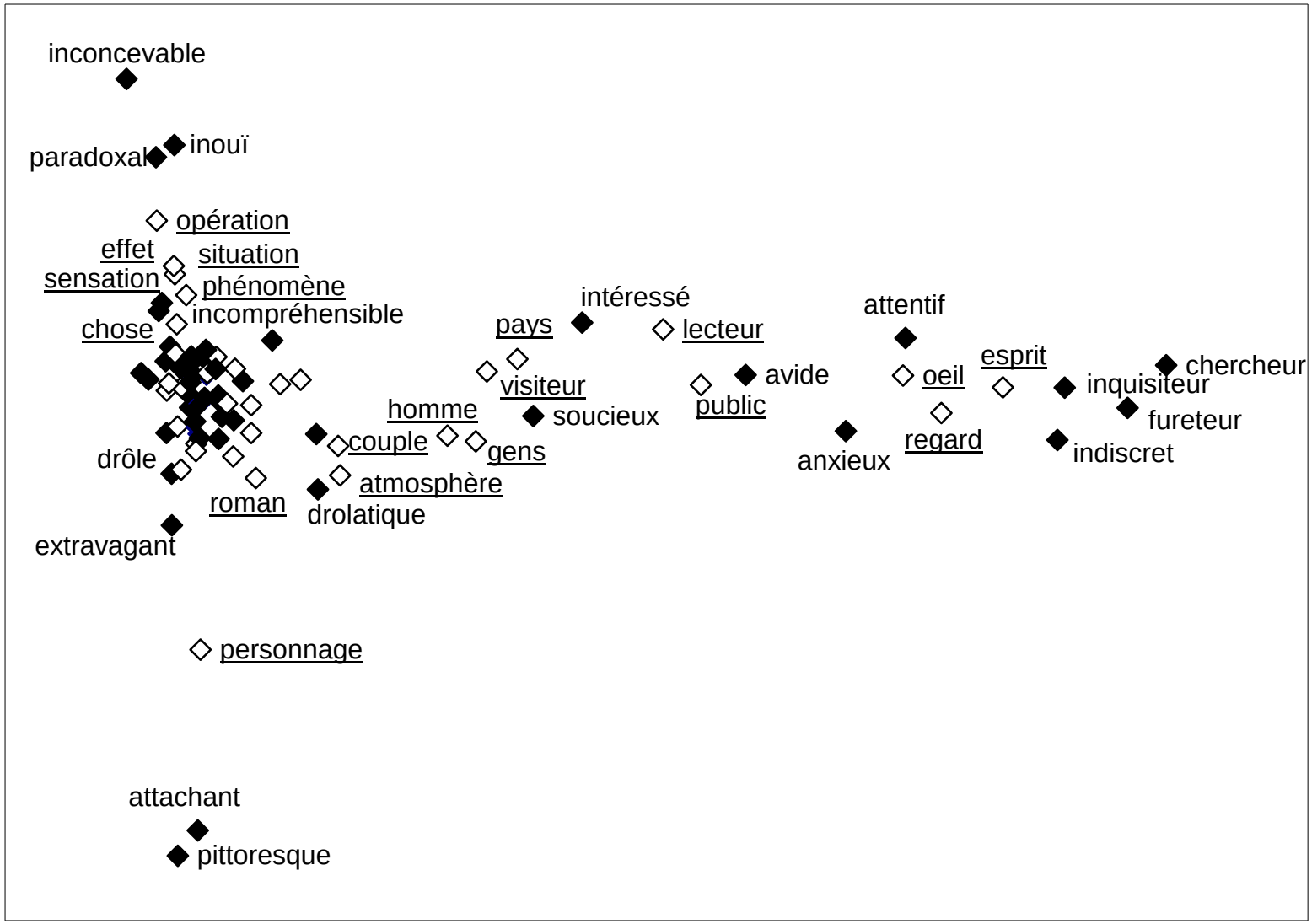


Figure 2 : espace sémantique obtenu avec les synonymes et les cooccurrences dans le corpus des périodiques

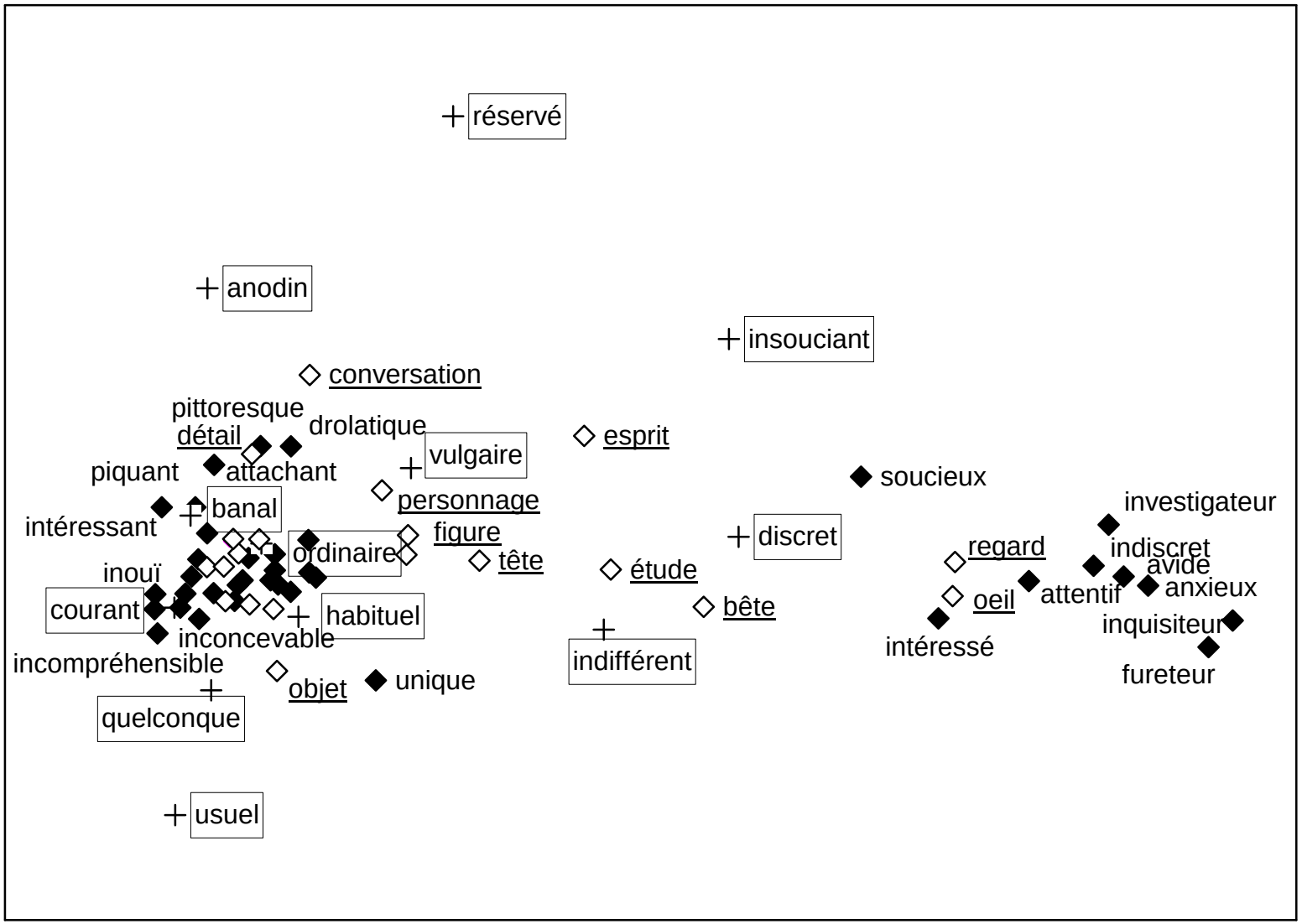


Figure 3 : espace sémantique obtenu avec les synonymes, les antonymes et les cooccurrences dans le corpus Frantext

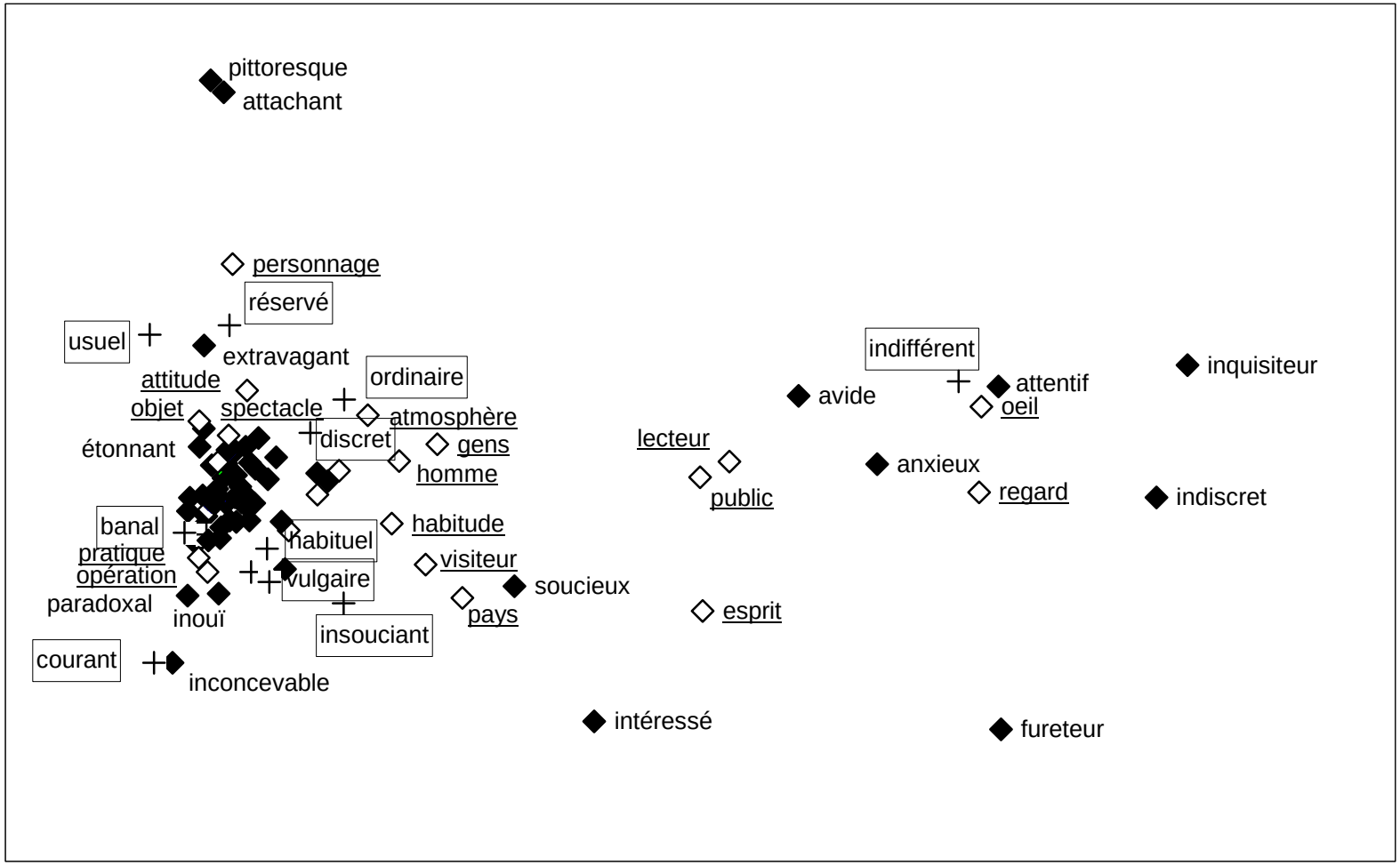


Figure 4 : espace sémantique obtenu avec les synonymes, antonymes et cooccurrences dans le corpus des périodiques

ORIENTATION MACHINE

UNE PLATE-FORME LOGICIELLE ET UNE DEMARCHE POUR LA VALIDATION DE RESSOURCES LINGUISTIQUES SUR CORPUS : APPLICATION A L'EVALUATION DE LA DETECTION AUTOMATIQUE DE CADRES TEMPORELS

Stéphane FERRARI[(1)], Frédérik BILHAUT[(1)],
Antoine WIDLÖCHER[(1)] et Marion LAIGNELET[(2)]
(1) GREYC – CNRS UMR 6072, Université de Caen – Basse-Normandie
(2) ERSS – CNRS UMR 5610, Université de Toulouse – Le Mirail

1. INTRODUCTION

L'orientation des travaux menés en Extraction d'Information, en Recherche d'Information ou concernant les systèmes de Question-Réponse multiplie et renforce les besoins en analyse de textes. En TAL, Traitement Automatique des Langues, de nombreuses approches visent désormais le repérage et la caractérisation de structures de textes (Enjalbert, 2005), en terme de mise en relation (Hernandez, 2004), de segmentation linéaire (Hearst, 1994) ou thématique (Bilhaut & Enjalbert, 2005). Ces approches à l'échelle du discours sont en rapport étroit avec des applications telles que le résumé automatique (Boguraev & Kennedy, 1997 ; Saggion & Lapalme, 2002) ou encore la navigation intra documentaire (Minel, 2002 ; Couto *et al*., 2004). De la même manière, des études en linguistique se fondant sur l'analyse de corpus se focalisent elles aussi sur cette dimension discursive afin de mettre en évidence des formes d'organisation textuelle. Elles s'inscrivent ainsi dans un courant de travaux de cette discipline et d'autres connexes qui s'appuient sur l'essor des ressources et des corpus électroniques (Williams, 2005). Dans cette optique, des outils informatiques peuvent servir pour l'exploration des corpus ainsi que pour l'expérimentation des modèles eux-mêmes en automatisant tout ou partie de ces processus (Péry-Woodley, 2005).

Les travaux que nous présentons ici s'inscrivent dans cette lignée, à l'intersection du TAL et de la linguistique. Initiés par le projet GéoSem[1], relatif au traitement sémantique de l'information géographique, ils ont pour objectif la mise au point d'une démarche et la réalisation des outils adéquats pour faciliter tant la validation de ressources linguistiques que l'évaluation des résultats de leur mise en œuvre sur corpus. Nous illustrons notre démarche à travers l'étude du problème de l'analyse des cadres temporels. Il s'agit d'un type particulier de cadres de discours tels qu'introduits par Charolles (1997). L'auteur a formulé l'hypothèse de l'existence de segments textuels d'une nature singulière, amenés par un adverbial en position saillante forte, appelé introducteur de cadre, et dont la sémantique s'étend à l'ensemble du segment textuel ainsi introduit. De tels cadres jouent un rôle important dans l'organisation du discours. Les cadres temporels en constituent une classe spécifique particulièrement pertinente pour l'analyse du discours en rapport avec

l'information géographique. Un analyseur automatique a été réalisé pour permettre la détection et l'annotation de ces structures textuelles (Bilhaut *et al*., 2003b). La première partie de l'article présente le modèle qui sous-tend cette réalisation, en insistant sur la plate-forme logicielle LinguaStream (Widlöcher & Bilhaut, 2005), un laboratoire virtuel d'expérimentation sur corpus qui a permis une mise en œuvre entièrement déclarative de cette approche. Nous précisons à quels besoins ses fonctionnalités répondent, en situant son utilisation dans un cycle d'expérimentation et en illustrant son exploitation dans la chaîne de traitements complexes permettant l'analyse de cadres temporels. LinguaStream constitue un élément de solution au problème de la validation de modèles par expérimentation sur corpus.

L'évaluation proprement dite pose quant à elle de multiples questions auxquelles nous tentons de répondre par l'élaboration d'un protocole relativement complexe. Des protocoles existent pour évaluer la qualité d'outils de TAL, mais leur adéquation à l'analyse de cadres de discours n'est pas immédiate. Les résultats d'une première expérience, exprimés en termes de précision et de rappel comme cela se pratique usuellement, permettent d'illustrer ce phénomène. L'évaluation d'une analyse automatique de cadres de discours ne peut évidemment se limiter à ces premières mesures. Nous en proposons d'autres afin de pallier les manques concernant notamment l'évaluation d'une propriété caractéristique du phénomène étudié : le positionnement des bornes des segments repérés. Les résultats concernant ce point mettent en évidence un comportement plutôt inattendu de l'analyseur automatique. Ils permettent selon nous de proposer de nouveaux éléments de solution au problème général de l'évaluation dans le cadre d'analyses de texte, en élaborant les prémisses d'un modèle d'évaluation qui prenne en considération l'indécision de l'expert humain dans la mesure de la qualité des résultats d'un traitement automatisé.

2. MODELE OPERATOIRE ET MISE EN ŒUVRE

Au risque de formuler une tautologie, il convient de rappeler la nécessaire traduction qu'impose la validation d'un modèle linguistique par expérimentation sur corpus. Projeter sur une collection de textes informatisés des connaissances relatives à un modèle linguistique ne peut s'opérer sans une reformulation de ce modèle en un langage formel qui permettra son traitement par un processeur. « Reformulation » paraît être un terme par trop neutre pour désigner un tel processus de formalisation ; « traduction » semble certes mieux approprié, bien qu'il ne couvre pas non plus l'étendue des transformations que peut subir un modèle pour devenir « computable ». C'est précisément le problème de la réécriture d'un modèle linguistique descriptif en un modèle opératoire que nous abordons dans cette partie. Nous montrons aussi en quoi le choix des outils informatiques constitue un point central dans pareil processus, leur expressivité déterminant les traductions possibles.

2.1. Formalisation d'un modèle

2.1.1. DU DESCRIPTIF AU PRESCRIPTIF L'hypothèse linguistique des cadres de discours que propose Charolles (1997) n'est pas directement un modèle « opératoire » pour une expérimentation sur corpus. Il s'agit plutôt d'un modèle à la fois descriptif et définitoire. L'auteur y présente un type d'unité textuelle dont la spécificité réside dans l'interprétation que nous en faisons dans le discours, interprétation subsumée par un critère sémantique stable. L'introducteur de cadre est l'élément du discours qui véhicule un tel

critère sémantique. Il s'agit d'une expression, en général un adverbial, en position détachée et à l'initiale d'une proposition, qui fonctionne comme une marque d'indexation, dans la mesure où il est capable de regrouper, d'indexer plusieurs propositions dans un fonctionnement « vers l'avant » (Charolles, 2002).

Selon le critère sémantique que porte son introducteur, un cadre de discours revêt une nature particulière : spatial, thématique, organisationnel, … Un cadre de discours temporel correspond au cas où l'introducteur de cadre est lui-même une expression temporelle. Le discours géographique est par exemple caractérisé par l'expression d'une information souvent située dans l'espace et dans le temps, comme l'illustre l'exemple de la figure 1.

De 1965 à 1985, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible *dans le Sud-Ouest et le Massif central*, modérée *en Bretagne et à Paris*, l'augmentation a été considérable *dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne* où les effectifs ont souvent plus que doublé.

Figure 1 : Un exemple de cadre temporel repris de (Bilhaut et al., 2003a). L'introducteur de cadre (en gras) est une expression temporelle ; les informations spatiales (italique) constituent une autre dimension caractéristique de l'information géographique.

L'hypothèse linguistique des cadres de discours temporels décrit certaines propriétés remarquables, les unes relatives à l'introducteur de cadre, les autres relatives aux cadres eux-mêmes. Pour les premières, il s'agit à la fois de caractéristiques propres telle que la nature syntagmatique des introducteurs, et de propriétés, notamment positionnelles, qui mettent en relation l'introducteur avec le discours dans lequel il apparaît. Les secondes concernent le contenu du cadre, parfois en relation avec le contexte discursif. Transformer le modèle linguistique en un modèle opératoire revient à ériger en contraintes ces caractéristiques pour permettre un repérage automatique des cadres. La démarche de validation sur corpus de l'hypothèse linguistique consiste bien ainsi à observer les éléments du texte qui correspondent à la description du modèle linguistique, de sorte à juger de leur pertinence. Il convient de noter que ce processus dénature le modèle initial, dans la mesure où il transforme des caractéristiques descriptives en conditions prescriptives. L'observation des éléments correspondant revient alors en quelque sorte à juger du caractère nécessaire et du caractère suffisant de ces conditions, notamment lorsqu'elle aboutit à une évaluation.

2.1.2. MODÈLE OPÉRATOIRE DES CADRES TEMPORELS Dans cette optique, le modèle opératoire des cadres a été décomposé en deux procédés successifs : l'un pour détecter les introducteurs potentiels, l'autre pour en déterminer la portée, c'est-à-dire délimiter les cadres proprement dits. Ces deux modèles n'ont pas été formalisés de la même manière. Bien que présenté comme une étape, détecter automatiquement les introducteurs temporels peut encore être vu comme une succession d'étapes, de traitements automatiques apportant chacun un nouvel élément d'information aussitôt exploitable pour l'étape suivante.

Ainsi, avant de statuer sur leur fonction d'introducteur, il faut tout d'abord procéder à un repérage des expressions temporelles. L'analyse sémantique de ces dernières a donné lieu à différents travaux tels que ceux présentés dans Bilhaut *et al*. (2003a). Opérant au niveau syntagmatique, et pouvant s'appuyer principalement sur des contraintes morphosyntaxiques, des ressources lexicales réduites et des patrons assez rigides, ce type

d'analyse vise une interprétation sémantique de l'expression et peut en outre fournir une représentation de l'intervalle temporel correspondant, que nous conservons pour son exploitation ultérieure par le modèle d'analyse des cadres. Déterminer le rôle d'introducteur des expressions repérées consiste alors à exploiter les caractéristiques contextuelles, c'est-à-dire ici la position détachée, à l'initiale de la proposition, comme autant de contraintes pour qu'une expression acquière un tel rôle.

Les introducteurs étant repérés, il reste à déterminer leur portée, c'est-à-dire délimiter les bornes des cadres temporels. L'introducteur étant déjà en position initiale, le problème qui demeure est la délimitation de la borne finale. Une étude menée par Le Draoulec & Péry-Woodley (2003) a décrit et étudié le comportement de différents marqueurs potentiels de fin de cadre, tels que le changement des temps verbaux et le changement de paragraphe. Ces marqueurs fonctionnent préférentiellement de manière conjointe, et il est difficile de dresser une liste exhaustive des divers mécanismes entrant en jeu. Ces propriétés permettent donc d'énoncer une série de critères qui constituent les conditions sur lesquelles fonder l'analyse automatique.

- Critères énonciatifs : à l'intérieur d'un même cadre, les temps verbaux devront être homogènes. Cependant, des variations de temps sont acceptables à l'intérieur de certains groupes compris dans le cadre, mais isolés, telles que des parenthèses et des subordonnées.
- Critères de cohérence sémantique : différents calculs sémantiques permettent de vérifier que l'intervalle temporel d'une expression temporelle présente dans le cadre est compris dans l'intervalle fixé par son introducteur.
- Critères structurels : un certain nombre de contraintes structurelles doivent être vérifiées. Tout d'abord, le cadre ne pourra être composé que de phrases complètes. Par ailleurs, une taille maximale est également fixée : si les différentes contraintes précédentes sont vérifiées (*i.e.* si aucune borne de fin n'a pu être déterminée) et qu'une fin de paragraphe est atteinte, on considère que cette dernière constitue la clôture recherchée.

Enfin, nous ignorons ici le problème des cadres imbriqués en considérant que la présence d'un nouvel introducteur impose la fermeture du cadre précédent. Cette contrainte et l'application systématique des critères précédents sont si rigides qu'ils laissent supposer une fermeture prématurée de certains cadres, hypothèse qui semble contredite par les premiers résultats de l'évaluation présentés plus loin.

2.2. Mise en œuvre informatique

2.2.1. DU PRESCRIPTIF A L'EFFECTIF Le modèle opératoire précédent ne constitue toujours pas l'étape finale des transformations que subit un modèle linguistique dans un processus de validation. Pour donner lieu à une réalisation informatique effective, il reste encore à l'exprimer en faisant un choix parmi les multiples formalismes existant, mais aussi parmi les outils logiciels qui les mettent en œuvre. Or ce dernier choix limite généralement l'utilisateur à un formalisme unique, alors qu'une expérimentation sur corpus nécessite souvent une chaîne de traitements multiples qui ne s'expriment pas nécessairement dans le même formalisme. La plate-forme LinguaStream2 a été conçue pour pallier ce problème. Elle propose un laboratoire virtuel pour l'expérimentation donnant accès à de multiples formalismes d'expression de contraintes linguistiques, tout en assurant une prise en main aisée par des non spécialistes. Nous renvoyons à Widlöcher &

Bilhaut (2005) pour une présentation des principes fondamentaux de LinguaStream et des formalismes qui y sont utilisables.

Les modèles opératoires précédents ont donné lieu à un ensemble de traitements exploitant chacun divers formalismes en reprenant et en détaillant les étapes d'analyses successives identifiées. La figure 2 présente la chaîne de composants qui symbolisent chacun de ces traitements, allant de la lecture du corpus à l'affichage des résultats.

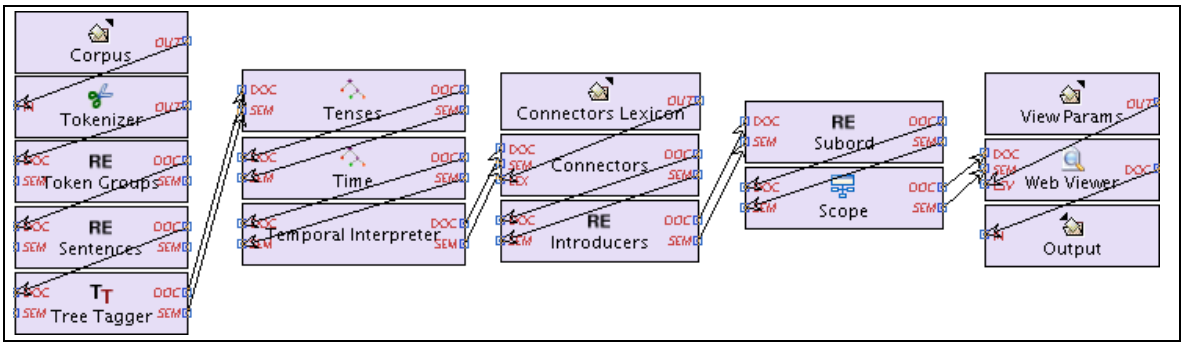


Figure 2 : La chaîne d'analyse des cadres temporels.
Dans LinguaStream, chaque module ou ressource d'une chaîne de traitement est représenté par une boîte, des connecteurs symbolisés par des flèches reliant les boîtes précisent quant à eux quelles informations les composants se transmettent.

Dans la partie gauche de la figure, les première étapes constituent des prétraitements dont les résultats conduisent au repérage et à l'annotation des expressions temporelles : segmentations successives en mots (*Tokenizer*), en groupes (*Token Groups*) puis en phrases (*Sentences*), étiquetage morphosyntaxique (*Tree Tagger*), analyse des temps des verbes (*Tenses*) par une grammaire DCG, analyse de l'expression du temps et construction des représentations correspondantes (*Time* et *Temporal Interpreter*). La figure 5 permet d'apprécier le type d'information associée à une expression temporelle (boîte 38). La suite de la chaîne permet un repérage d'éventuels connecteurs discursifs (mais, cependant, etc.), puis l'analyse des introducteurs. Les critères positionnels (*RE – Introducers*) sont exprimés à l'aide du formalisme MRE (*Macro Regular Expressions*) proposé par LinguaStream pour l'écriture de transducteurs pouvant s'appliquer autant au texte qu'à des unités issues d'analyses précédentes. L'exemple de la figure 3 utilise ce formalisme déclaratif, permettant une écriture de règles relativement proche de la manière dont les contraintes du modèle peuvent être exprimées en langue naturelle.

```
{type:phrase, anchor:start}
<introducteur>
{type:connecteur}? {type:temporel} /as $t
</introducteur> /sem {axe:temps, valeur:$t}
","
```

Figure 3 : Un exemple de contrainte de l'analyseur des introducteurs de cadres temporels.
La règle présentée exprime notamment le fait qu'un introducteur doit être une expression de type temporel, au début d'une phrase, éventuellement précédée d'un connecteur.

L'analyse de la portée des cadres (composant *Scope*) s'appuie quant à elle sur un autre formalisme déclaratif dédié à l'analyse du discours, issu des travaux de Widlöcher (2004) et présenté dans Widlöcher & Bilhaut (2005), permettant l'expression de contraintes non linéaires (sur des éléments potentiellement distants) et non séquentielles (l'ordre de ces

éléments pouvant n'être pas spécifié) conformément à un besoin selon nous récurrent pour l'analyse de structures discursives. Un exemple de contrainte est fourni dans la figure 4. Le résultat de ces analyses est consultable dans un navigateur classique, comme l'illustre la figure 5.

```
Rule {type:"cadre"}:
        start(pattern:{type:"introducteur"})
        end(pattern:{type:"phrase"})
        homogeneity(comparator:portée)
        size(mode:#LONGEST)
Comparator portée ({type:"verbe"} as $v1, {type:"verbe"} as $v2):
        $v1/temps = $v2/temps
Comparator portée ({type:"introducteur"} as $i, {type:"temporel"} as
$t):
        ($t/debut >= $i/debut) and ($t/fin <= $i/fin)
```

Figure 4 : Un exemple de contrainte de l'analyseur de portée.
La règle désigne un élément commençant par un introducteur, composé de phrases complètes, dont tous les verbes sont au même temps, et dans lequel les expressions temporelles sont compatibles avec l'intervalle fixé par l'introducteur.

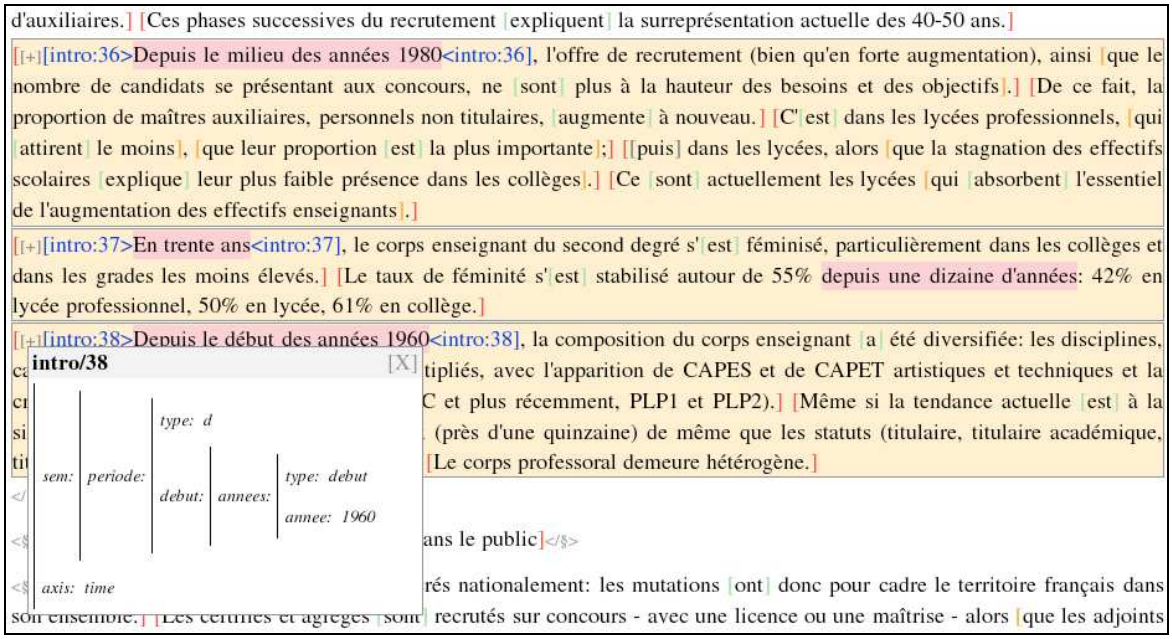


Figure 5 : Un exemple de sortie de l'analyseur de cadres temporels.
Trois cadres successifs et leurs introducteurs sont mis en valeurs, précédés chacun d'une croix interactive ouvrant une « boîte » contenant la structure de traits associée.

LinguaStream permet aussi une consultation ciblée sur un type d'élément annoté dans le corpus à l'aide de concordanciers, classique ou « macro » pour l'observation d'annotations concernant des éléments de discours de grande taille dont le texte n'est plus affiché dans sa totalité. La plate-forme exploite ici les principes observés par Ferrari et Perlerin (2003). Elle permet en outre d'intégrer des composants extérieurs (c'est le cas du *Tree Tagger* dans l'exemple présenté) ou de réutiliser des paquets de composants sous forme de macro.

La validation de modèles et de ressources linguistiques sur corpus nécessite donc un processus complexe de traduction auquel nous apportons des éléments de solutions avec la plate-forme LinguaStream qui offre à la fois de nombreuses possibilités d'expression selon

différents formalismes et une interface graphique pour enchaîner les composants ainsi décrits. À ce stade de notre étude du processus d'expérimentation sur corpus, LinguaStream permet donc la mise en œuvre d'analyses automatiques pour la détection et l'annotation d'éléments textuels, ainsi que l'observation des résultats de ces analyses. Dans la partie suivante, nous abordons le problème de l'évaluation de tels résultats, phase finale et pour ainsi dire incontournable de l'expérimentation.

3. PROBLEMATIQUE DE L'EVALUATION

En étudiant les cadres temporels, il apparaît en définitive de nombreuses motivations pour mener non pas une, mais plusieurs évaluations différentes, qu'il est possible de classer selon la typologie suivante :

1) évaluations intrinsèques de la qualité de la détection des introducteurs, de la pertinence des annotations sémantiques qui leur sont associées, de la pertinence de la portée des cadres de discours annotés ;
2) évaluations extrinsèques, s'orientant plutôt vers une analyse de la valeur ajoutée de l'annotation de cadres temporels pour des applications tierces.

Des évaluations extrinsèques sont menées dans le cadre du projet GéoSem avec des partenaires à l'Université de Pau et des Pays de l'Adour ainsi qu'à l'Université de Caen – Basse-Normandie. Leur objectif est de mesurer la valeur ajoutée par l'application finale dans des tâches précises de recherche d'information géographique. Nous nous focalisons pour notre part sur les problèmes déjà nombreux liés à une évaluation intrinsèque. Une première expérience a été réalisée à l'aide d'un corpus de référence : un expert a proposé une annotation des cadres de discours temporels sur un ouvrage de géographie[3]. Les mesures présentées dans cette partie en sont issues. Une discussion est proposée en fin de partie pour proposer un protocole plus complexe et mieux adapté à l'ensemble des éléments à prendre en considération pour mener l'évaluation.

3.1. Précision, rappel et mesures relatives

Les techniques d'évaluation d'outils de TAL s'appuient classiquement sur des mesures particulièrement bien adaptées à des problèmes de détection : la précision et le rappel, ou les dérivées que sont le bruit, le silence et la F-mesure. Nous interrogeons leur adéquation non seulement à l'évaluation d'un outil automatique tel l'analyseur de cadres temporels issu des travaux présentés dans la partie précédente mais aussi au problème plus générique de l'expérimentation sur corpus d'un modèle linguistique. Il s'agit donc moins d'évaluer un outil de TAL en tant que tel que d'étudier les retombées de cette évaluation sur le modèle linguistique initial et le modèle opératoire dérivé.

3.1.1. UN BRUIT imprécis La précision consiste à évaluer la proportion de vrais cadres temporels parmi l'ensemble des unités repérées. Le bruit est la proportion complémentaire, celle de mauvaises réponses. Il s'agit pour un outil de TAL d'une appréciation quantitative de sa qualité. Mais dans un processus expérimental tel que celui qui nous intéresse, le bruit permet avant tout d'apprécier dans quelle mesure les propriétés énoncées sont caractéristiques du phénomène étudié : il représente la proportion d'éléments textuels détectés par les critères mais ne relevant pas a priori du phénomène attendu. Un bruit élevé reste donc à interpréter au regard de la volonté initiale d'être exhaustif dans la caractérisation du phénomène, qui n'est pas nécessairement le but poursuivi.

Dans l'expérience que nous avons menée, l'analyse automatique a permis de détecter 80 cadres. Parmi ceux-ci, 74 ont pu être alignés avec les annotations de l'expert, la précision relative est 92,5%, le bruit de 7,5%. En analysant les éléments détectés automatiquement mais ne correspondant pas à une annotation de l'expert, il ressort 3 oublis de l'expert, 2 problèmes de protocole, et une seule erreur due à un mauvais type d'expression temporelle. Les problèmes de protocole correspondent à des cadres imbriqués, pour lesquels l'expert n'a annoté que le cadre englobant, que l'analyseur n'a pas repéré. Les oublis de l'expert font l'objet d'une discussion ci-après. En définitive, un seul des éléments ne correspond pas à un cadre, et ce bruit est à imputer à une erreur d'un module d'analyse en amont de ceux dédiés aux cadres. Ce résultat relativise grandement le bruit calculé, mesure imprécise parce que peu en rapport avec le phénomène initialement étudié.

3.1.2. UN RAPPEL À PASSER SOUS SILENCE Le rappel permet de mesurer le nombre d'unités correctes détectées par rapport au nombre total d'unités du type recherché dans l'ensemble du corpus. Le silence est le complémentaire, la proportion d'unités oubliées. Ici encore, il s'agit en TAL d'une appréciation quantitative de la qualité d'une détection. Dans notre cas, un silence élevé peut témoigner de certains écueils, comme une caractérisation trop précise délimitant une sous-classe du phénomène, mais il peut aussi ne s'agir que de ressources limitées à certains niveaux des analyses, l'expérimentation ne visant pas nécessairement l'exhaustivité sur ce point non plus. Enfin, qu'il s'agisse de rappel ou de précision, les limitations peuvent toujours être dues à la mise en œuvre, l'utilisation de chaînes de composants multipliant les sources d'erreurs ou d'oublis.

L'expert a trouvé 156 cadres, le rappel correspondant est donc 47,4%, le silence de 52,6%. L'analyse de quelques exemples de cadres oubliés montre des comportements réguliers qui permettent d'expliquer le silence obtenu. L'analyseur d'expressions temporelles utilisé se limite à celles qui sont explicitement datées. Or, de nombreux cadres parmi ceux non trouvés ont la particularité de contenir des introducteurs dont l'aspect temporel est peut-être immédiat pour un interprétant humain, mais difficilement calculable. Ainsi, dans *« À l'occasion du mouvement national … »*, il est fait référence à un événement, dont nous pouvons d'ailleurs ignorer la date précise en contexte. De telles situations restent donc difficiles à traiter, et certaines ne peuvent l'être dans la mesure où ce sont souvent des connaissances extérieures qui seraient nécessaires pour dater l'événement au final. D'autres cadres oubliés ne correspondent pas tout à fait au modèle linguistique ayant servi de référence. Il s'agit par exemple d'introducteurs qui ne sont pas détachés : *« Pour la fin des années 1980 à défaut de statistiques… »*. Il s'agit donc de pondérer les silences qui ne sont pas tous imputables au composant analyseur de cadres temporels proprement dit, mais à des étapes préalables de la chaîne de traitement.

3.2. Compléments mesurés

Si les mesures précédentes ne paraissent pas très adaptées à l'expérimentation menée, il en est d'autres, relativement simples à mettre en œuvre, qui le sont beaucoup plus. Ainsi, la difficulté ressentie lors de l'élaboration de tous les modèles concernant la détermination de la borne droite des cadres invite à observer et quantifier ce phénomène précis. Nous avons de ce fait mesuré quelques longueurs et écarts en nombre de mots. La taille moyenne des cadres de l'expert est de 120 mots, tandis que celle des cadres détectés automatiquement est de 175 mots. Plus précisément, il s'avère que les bornes gauches des cadres alignés sont toutes confondues, tandis que la borne droite d'un cadre automatiquement détecté est confondue ou déborde systématiquement de celle posée par l'expert, avec un écart moyen

de 48 mots. Ces quelques nombres apportent beaucoup plus d'information pour la validation du modèle que les mesures précédentes. Le dépassement systématique, comportement imprévu étant donnée la rigueur des règles, reflète en grande partie une propriété inhérente au phénomène étudié. En situation d'expertise, l'interprétation d'un cadre selon le critère sémantique de l'introducteur conduit à repérer des zones que nous qualifions de floues, et pour lesquelles le critère peut s'appliquer, mais sans s'imposer nécessairement. En réaction, l'expert semble avoir fermé de tels cadres au début de ces zones plutôt qu'à leur fin, quand l'analyseur les prolonge au contraire tant que les contraintes sont vérifiées.

L'analyse d'éléments du discours s'articule de plus en plus sur des unités textuelles de grande taille dont la détermination précise pose problème. Le besoin émerge donc d'un protocole d'évaluation adapté à ce type d'analyses, qui permette de tenir compte du flou, en plus de la variabilité entre humains. La poursuite des travaux relatifs aux cadres de discours temporels va dans ce sens, avec la proposition d'un modèle d'annotation autorisant plusieurs bornes de fin, et la constitution d'un corpus de référence ainsi annoté par plusieurs experts afin d'apprécier la variabilité de l'ensemble des mesures. Les premiers résultats permettent déjà de relativiser les rares oublis constatés dans l'expertise précédente, 3 sur 156, une grande variabilité, proche de 45%, ayant été observée sur un premier échantillon. Ce sont principalement les types de cadres temporels annotés qui varient selon les experts : cadres introduits par un expression temporelle effectivement datée, cadres introduits dans les titres… Les résultats de cette expérimentation et la présentation du protocole qui y est associé fera l'objet d'une autre publication.

4. MISE EN PERSPECTIVE CONCLUSIVE

Les travaux présentés relatifs à l'analyse des cadres de discours temporels ont permis de mettre en évidence la multiplicité des transformations nécessaires à la réécriture d'un modèle linguistique pour sa validation par expérimentation sur corpus et la nécessité d'élaborer des protocoles d'évaluation adaptés aux phénomènes étudiées en analyse du discours. En ce qui concerne l'expérimentation d'un modèle descriptif, sa traduction en un modèle opératoire puis la mise en œuvre informatique de ce dernier constituent des étapes au cours desquelles une dégradation est possible. Afin d'y pallier, la plate-forme LinguaStream a été développée. Elle multiplie les formalismes pour éviter tout dictat en ce qui concerne le choix de la mise en œuvre. Elle offre des fonctionnalités de visualisation et de navigation dans le corpus adaptées à l'expérimentation de modèles linguistiques.

Le problème de l'évaluation se pose encore, une fois résolu celui de l'expérience. L'analyse des résultats d'une confrontation entre annotations automatiques et annotations d'expert a mis en évidence le manque de pertinence des mesures prévues pour apprécier la qualité d'outils de TAL. Sans contester leur intérêt, elles ne permettent pas un retour immédiat sur le modèle initial. À l'inverse, des mesures de taille et d'écart permettent un tel retour. Cependant, une notion de flou semble caractéristique des unités textuelles étudiées, et peut-être plus généralement de celles étudiées en analyse du discours. La nécessité s'impose donc de proposer de nouveaux protocoles d'évaluation qui prennent ce phénomène en considération.

5. NOTES DE FIN

1. GéoSem : « traitement sémantique de l'information géographique », programme CNRS « société de l'information ».
2. http://www.linguastream.org
3. *Atlas de la France scolaire de la maternelle au lycée*, R. Hérin.

6. REFERENCES

Bilhaut F., Charnois T., Enjalbert P. & Mathet Y. 2003a. 'Passage extraction in geographical documents'. In *Proceedings of New Trends in Intelligent Information Processing and Web Mining*, Zakopane, Poland.

Bilhaut F. & Enjalbert P. (2005). Discourse thematic organisation reveals domain knowledge structure. In Proceedings of the 2nd Indian International Conference on Artificial Intelligence, Pune, India.

Bilhaut F., Ho-Dac M., Borillo A., Charnois T., Enjalbert P., Le Draoulec A., Mathet Y., Miguet H., Péry-Woodley M.-P. & Sarda L. 2003b. 'Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique'. In B. Daille, Ed., *Actes de TALN 2003* (Traitement automatique des langues naturelles), Batz-sur-Mer, France : ATALA IRIN.

Boguraev B. & Kennedy C. 1997. 'Salience-based content characterisation of text documents'. In *Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization*, p. 2–9, Madrid, Spain.

Charolles M. 1997. 'L'encadrement du discours : Univers, champs, domaines et espaces'. *Cahier de Recherche Linguistique* n°6. Université de Nancy 2.

Charolles M. 2002. *Les adverbiaux cadratifs : fonction et classification*.

Couto J., Ferret B., Grau B., Hernandez N., Jackiewicz A., Minel J. & Porhiel S. (2004). 'Régal, un système pour la visualisation sélective de documents'. *La présentation d'information sur mesure, RIA*, Numéro spécial, 18(4), 481–514.

Enjalbert P. (dir.) 2005. *Sémantique et traitement automatique du langage naturel***.** Traité IC2, Hermès. Paris : Lavoisier**.**

Ferrari S. & Perlerin V., 2003. 'Les besoins d'interactions en traitement automatique des langues et en linguistique de corpus : étude de cas'. *Actes des 3èmes Journées de la Linguistique de Corpus*, Lorient, France. À paraître aux Presses Universitaires de Rennes.

Hearst M. 1994. 'Multi-paragraph segmentation of expository texts'. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.

Hernandez N. 2004. *Description et Détection Automatique de Structures de Texte.* Thèse de doctorat de l'Université de Paris XI.

Le Draoulec A. & Péry-Woodley M.-P. 2003. 'Time travel in text : Temporal framing in narratives and non-narratives'. ERSS, Université de Toulouse-Le Mirail.

Minel J.-L. 2002. *Filtrage sémantique, du résumé automatique à la fouille de textes*. Hermès.

Péry-Woodley M. 2005. 'Discours, corpus, traitements automatiques', In A. Condamines, Ed., *Sémantique et Corpus*. Hermès.

Saggion H. & Lapalme G. 2002. 'Generating informative and indicative summaries with SumUM'. *Computational Linguistics*, 28(4), 497–526.

Schmid H. 1994, 'Probabilistic Part-of-Speech Tagging Using Decision Trees', *International Conference on New Methods in Language Processing*. Manchester, UK.

Widlöcher A. 2004. Analyse macro-sémantique : vers une analyse rhétorique du discours. In *Actes de RECITAL 2004*, pages 183-188, Fès, Maroc.

Widlöcher A. & Bilhaut F. 2005. 'La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus'. In M. Jardino, Ed., *Actes de la 12e Conférence Traitement Automatique du Langage Naturel* (TALN), Dourdan, France : ATALA LIMSI.

Williams G. (ed.) 2005. *La linguistique de corpus***.** Rennes : Presses Universitaires de Rennes**.**

ALIGNEMENT D'UNITES TEXTUELLES DE TAILLE VARIABLE

Emmanuel Giguet (1), Marianna Apidianaki (2)
GREYC – CNRS UMR 6072– Université de Caen – Bd du Maréchal Juin – F14032 Caen Cedex Emmanuel.Giguet@info.unicaen.fr (2) Lattice – CNRS / ENS / Université Paris 7 – Denis Diderot Ecole Normale Supérieure, 1 rue Maurice Arnoux, F-92120 Montrouge Marianna.Apidianaki@linguist.jussieu.fr

1. INTRODUCTION

1.1. Traitement automatique de corpus bilingues et multilingues

Le traitement de corpus bilingues et multilingues parallèles, constitue un champ d'investigation très animé dans le domaine du traitement automatique des langues. Ces corpus sont constitués d'un ensemble de textes et de leurs traductions dans une ou plusieurs langues. Les informations linguistiques et traductionnelles qui peuvent être mises à jour par l'investigation et l'analyse de ces corpus les rendent une ressource importante tant pour les traducteurs, les lexicographes et les terminologues que pour les linguistes qui utilisent des méthodes empiriques pour l'étude contrastive des langues. Une autre raison qui a nourri l'intérêt envers ce type de ressources est qu'elles constituent le noyau des Systèmes de Traduction Basés sur l'Exemple et des Systèmes de Mémoires de Traduction.

Aujourd'hui, ces corpus sont plus aisément accessibles. De nouveaux sont proposés régulièrement par la communauté, que ce soit pour de nouvelles paires de langues que pour des thématiques spécifiques. Ils mettent à l'épreuve les modèles tant dans leur capacité à appréhender des langues variées qu'à s'adapter à des lexiques variés. L'augmentation de la taille de ces corpus pousse quant à elle vers la recherche de modèles opérationnels efficaces ne demandant que peu de supervision, tout en produisant de bons résultats, que ce soit du point de vue quantitatif ou bien qualitatif, et cela, dans le but de faciliter la description, l'investigation et l'analyse des informations contenues dans les textes.

1.2. Méthodes d'alignement de corpus parallèles

Un sujet « épineux » mais crucial pour toute la recherche tournant autour des corpus bilingues et multilingues est l'alignement, c'est-à-dire la localisation et l'extraction des relations de traduction existant entre les diverses unités constitutives des textes parallèles. Les méthodes d'alignement actuelles se situent sur un continuum allant des méthodes purement statistiques à celles intégrant des connaissances linguistiques détaillées. La nature des connaissances n'est en fait pas la seule manière de classer ces méthodes. Un angle tout aussi pertinent pour les étudier concerne la granularité des alignements proposés : la méthode propose-t-elle des équivalences de paragraphes, de phrases, de propositions, de termes, ou bien de mots, un choix qui dépend en fait des applications envisagées.

Pour l'alignement à gros grain (paragraphe, phrase), les principaux indices utilisés sont la longueur relative des segments, la présence de formes invariantes (certaines

ponctuations, certains symboles, nombres, sigles, etc.) et la mise en forme. Certaines méthodes vont dès ce stade introduire des dictionnaires bilingues. Ainsi, au « grain » phrase, Gale & Church (1991b), Brown *et al.* (1991) et Kay & Röscheisen (1993) proposent des méthodes d'alignement statistiques ; Simard *et al.* (1992) et Papageorgiou *et al.* (1994) ont, eux, recours aussi à des informations linguistiques.

À granularité fine, à savoir sous-phrastique, les méthodes statistiques se sont rapidement vues combiner à des descriptions linguistiques enrichies *via*, par exemple, un étiquetage morpho-syntaxique (tagging) ou une segmentation en groupes minimaux (chunking). Ainsi, on trouve des méthodes statistiques chez Gale & Church (1991a), Kitamura & Matsumoto (1995) et d'autres combinant des informations linguistiques (Tiedemann, 1993 ; Boutsis & Piperidis, 1996 ; Piperidis *et al.*, 1997). On notera également l'existence de méthodes proposant des alignements à un grain intermédiaire entre la phrase et le mot isolé (Smadja, 1992 ; Smadja *et al.*, 1996 ; Kupiec, 1993 ; Kumano & Hirakawa, 1994 ; Boutsis & Piperidis, 1998).

1.3. Limite des méthodes d'alignement

Le problème de l'alignement est par définition celui de la localisation et de la délimitation précise des segments à mettre en correspondance entre les langues. Même si l'alignement au grain paragraphe ou phrase semble plus accessible qu'un alignement sous-phrastique, on y rencontre aussi des difficultés. Une des faiblesses tient entre autre au fait que les méthodes d'alignement misent sur le parallélisme de la structure des documents et ne sont par conséquent que très peu tolérantes aux variations dispositionnelles du contenu.

Alors que l'alignement à gros grain s'appuie à la fois sur le marquage fort des paragraphes (*via* la mise en page) et faible des phrases (*via* la ponctuation), en corrélation avec le parallélisme présumé de la narration, l'alignement sous-phrastique se heurte immédiatement à la délimitation des unités, notamment lorsque le mot n'est pas physiquement marqué, ou bien lorsque la langue est agglutinante. À un grain sous-phrastique, on ne peut par ailleurs présumer une quelconque préservation de l'ordre des unités dans la phrase, sauf à se limiter à l'analyse de couples de langues proches. Pour pallier cette difficulté, le recours à un dictionnaire bilingue est souvent proposé, mais on peut regretter que les alignements produits ne soient alors que la mise en évidence de traductions déjà contenues dans le dictionnaire. On regrettera également que cette technique exclue l'analyse des langues faiblement dotées en matière de ressources linguistiques.

Enfin, on notera que l'alignement au niveau sous-phrastique est généralement effectué successivement à un alignement phrastique et que la qualité de ce dernier n'est donc pas sans effet sur la qualité des résultats du premier.

2. UNE APPROCHE ENDOGENE DE L'ALIGNEMENT SOUS-PHRASTIQUE

L'approche proposée ici ne prétend pas répondre à l'ensemble des limites précédemment énoncées. Elle concerne le repérage d'équivalents de traduction de taille variable, et notamment supérieure à un mot. La méthode est appliquée à un corpus parallèle de bi-textes anglais-grec. Nous n'adoptons pas de contraintes très rigides concernant la taille des unités linguistiques afin de tenir compte de la souplesse des langues et des divergences de traduction. Les liens d'alignement peuvent alors être établis à des niveaux variés – allant

de la phrase jusqu'au mot – en n'obéissant à d'autres contraintes qu'au nombre d'occurrences minimal. En ce qui concerne le prétraitement des textes d'entrée, ils ont été segmentés et alignés au niveau des phrases, mais ils ne portent aucune information morphosyntaxique et ne sont pas lemmatisés. Les divergences de flexion des mots isolés sont gérées automatiquement par un module intégré dans l'outil d'alignement, qui prend en compte les variants morphologiques. Ceci est très important surtout pour des langues comme le grec, qui est une langue hautement flexionnelle, avec une morphologie riche, et offre la possibilité de prendre en compte les formes fléchies différentes des mots.

2.1. Approche

Pour traiter ce problème, nous avons, dans la mesure du possible, essayé de nous tenir à une approche de type minimaliste, selon la méthode du groupe Syntaxe et Rhétorique de l'équipe ISLanD du GREYC. Nous n'utilisons donc pas de connaissances linguistiques particulières sur les langues à aligner.

En effet, beaucoup de langues sont peu dotées en matière de ressources libres pour le traitement automatique des langues, que ce soit par exemple en matière d'étiquetage morpho-syntaxique ou, plus généralement, d'analyse syntaxique de surface, ou bien tout simplement de ressources lexicales (dictionnaires électroniques etc.). De même, la possibilité de traitement de textes qui n'ont pas subi une longue chaîne de prétraitements est avantageuse vu la disponibilité de plus en plus grande de corpus bilingues.

Nous préférons donc des solutions fondées sur la recherche de motifs situés à des positions bien identifiées, combinée avec des indices de mise en forme matérielle (graisse, taille, couleur des caractères, centrage, justification etc.) ou bien typo-dispositionnelle (ponctuations, énumérations etc.).

2.2. Le corpus

Le corpus utilisé pour nos expériences est constitué de bi-textes alignés au grain phrase. Formellement, un bi-texte est un quadruplet <T1, T2, Fs, C> où T1 et T2 sont les deux textes, Fs est la fonction qui réduit T1 à un ensemble d'éléments Fs(T1) et T2 à un ensemble d'éléments Fs(T2), et C est un sous-ensemble du produit cartésien de Fs(T1) x Fs(T2) (Harris, 1988). Concrètement, un bi-texte aligné est constitué d'un ensemble d'alignements, c'est-à-dire un ensemble de correspondances entre phrases de la langue source et phrases de la langue cible. Ces alignements contiennent de 0 à 2 phrases par langue. Par exemple, un alignement « 2-1 » met en correspondance 2 phrases du texte 1 avec 1 phrase du texte 2, et un alignement « 1-0 » indique qu'une phrase du texte 1 n'a pas de correspondance dans le texte 2.

Le corpus est composé de 93 bi-textes anglais-grec (un peu plus de 200 000 mots). Il s'agit de résumés d'articles scientifiques parus dans des journaux internationaux de médecine (en anglais ou en grec) ainsi que leur traduction. Une partie du corpus est constituée de résumés d'articles parus dans le journal médical grec *Archives of Hellenic Medicine* de 1999 à 2004 (http://www.mednet.gr/archives/) et de leurs traductions en anglais. L'autre partie du corpus est constituée de résumés d'articles en anglais parus dans des journaux internationaux disponibles sur la base de données PubMed (www.pubmed.com). Les traductions de ces textes en grec ont paru dans des journaux publiés chez l'éditeur Scientific Publications : *Review of Risk Factors in Cardiology*, *Review of Diabetes*, *Review of Endocrinology – Metabolism*, *Review of Vaccines*, *Review of Paediatric Nutrition*. Il faut noter que même s'il s'agit de textes d'un domaine spécialisé

(médecine), il n'y a pas de grande homogénéité entre eux (surtout au niveau terminologique), étant donné qu'ils relèvent de sous-domaines variés (cardiologie, endocrinologie, pédiatrie, diabétologie, vaccinologie).

3. METHODE D'ALIGNEMENT AUTOMATIQUE

La méthode de résolution se déroule en deux phases basées sur deux hypothèses sous-jacentes. La première porte sur les documents et la seconde sur l'ensemble du corpus.

3.1. Hypothèses

• **Hypothèse 1** : soit un bi-texte composé des textes T1 et T2. Si une séquence S1 est répétée dans T1 dans un certain nombre de phrases P1i, nous supposons qu'une séquence S2 correspondant à la traduction de S1 apparaîtra dans les phrases P2j de T2 où P2j sont les phrases alignées avec P1i.
• **Hypothèse 2** : soit un corpus de bi-textes composé des langues L1 et L2. Il n'y a pas de garantie pour qu'une séquence S1 répétée dans plusieurs textes de L1 ait une unique traduction dans les textes correspondants de la langue L2.

3.2. Phase 1 : Alignement intra-document

3.2.1. PREPARATION DU CORPUS L'algorithme prend en entrée le corpus de bi-textes alignés au niveau phrastique. Cet alignement a été obtenu automatiquement, par l'outil d'alignement de phrases du Système de Mémoire de Traduction TrAid (Triantafyllou et al., 2000). L'algorithme utilisé se base sur le modèle statistique de Gale & Church (1991a) qui tient compte de la longueur des phrases en nombre de caractères. Le résultat du processus d'alignement a été validé à la main. La segmentation en phrases et leur alignement ont été les seuls prétraitements du corpus.

Le corpus aligné est au format TMX[1], codage UTF-8. Chaque alignement est balisé par un élément XML, qui contient les segments de chaque langue mis en correspondance pendant le processus d'alignement qui, eux aussi, sont inclus entre des balises XML (Figure 1).

```
<tu>
    <prop type="Domain">ALIGNER: [med1_EN.txt]---[med1_EL.txt]</prop>
    <tuv lang="EN">
        <seg>Frequency and relevance of elevated calcitonin levels in patients with neoplastic and nonneoplastic
        thyroid disease and in healthy subjects.</seg>
    </tuv>
    <tuv lang="EL">
        <seg>Συχνότητα και σημασία των αυξημένων επιπέδων καλσιτονίνης σε ασθενείς με και χωρίς
        νεοπλασματική νόσο του θυρεοειδούς και σε υγιή άτομα</seg>
        1 Translation Memory eXchange (http://www.lisa.org/tmx)

    </tuv>
</tu>
```

Figure 1. La structure des fichiers TMX

La première étape consiste à écarter du traitement les alignements des bi-textes pour lesquels des segments d'une langue n'ont pas de correspondance dans l'autre (alignement « 1-0 », « 0-1 », « 2-0 » et « 0-2 »). On profite de cette vérification pour relever dans les deux parties du corpus, les index (offsets) de début et de fin de chaque segment.

3.2.2. IDENTIFICATION DES SEQUENCES CANDIDATES A L'ALIGNEMENT Pour chacun des deux textes des bi-textes, nous calculons ensuite les séquences de mots répétées, ainsi que leur effectif (nombre d'occurrences). L'algorithme est paramétré par le nombre de mots minimal et maximal des séquences à mémoriser ainsi que par leur effectif minimal, mais ces paramètres ne sont utilisés qu'en phase de test. Par ailleurs, l'algorithme ne conserve pas les sous-séquences d'une séquence répétée si elles ont même effectif. Par exemple, si « heart disease » a même effectif que « coronary heart disease », nous ne retenons que la deuxième séquence. Par contre, si « disease » a un effectif supérieur à « coronary heart disease», nous retenons les deux.

Le calcul de l'effectif d'une séquence répétée s'accompagne du relevé des index de ses occurrences. On obtient donc à la fin du traitement une liste de séquences avec pour chacune son effectif, ainsi que la liste des index de ses occurrences dans le texte.

« healthy subjects » : index des 22 occurrences dans le texte 1 (44 segments)

285, 531, 984, 2860, 3993, 4281, 4386, 4646, 4781, 5122, 5242, 5382, 5473, 6615, 7370, 7422, 7690, 8856, 9019, 9459, 9553, 10025

3.2.3. PASSAGE À UNE REPRÉSENTATION VECTORIELLE DES SÉQUENCES Nous construisons ensuite un espace orthonormé pour explorer l'existence de relation de traduction entre séquences et définir les couples traductionnels internes à chaque bi-texte. Pour cela, nous notons pour chaque occurrence de séquence répétée les numéros où elle apparaît.

« healthy subjects » : liste des alignements où apparaît la séquence

2 , 4 , 7 , 13 , 15 , 16 , 16 , 17 , 17 , 18 , 19 , 20 , 21 , 30 , 34 , 34 , 35 , 39 , 39 , 41 , 41 , 42

Nous convertissons cette liste en un vecteur à *n* dimensions (où *n* correspond au nombre d'alignement du bi-texte). Chaque dimension contient le nombre d'occurrences de la séquence présent dans l'alignement.

« healthy subjects » : vecteur à 44 dimensions associé

1234567891011121314151617181 9…4142434401010010000010122 11…2100

3.2.4. ALIGNEMENT DES SÉQUENCES À L'AIDE DU COSINUS Nous distinguons dorénavant les deux langues du corpus, la langue L1 et la langue L2. Pour chaque séquence candidate à l'alignement dans la langue L1, nous explorons l'existence d'une relation de traduction entre elle et chacune des séquences de la langue L2 candidates à l'alignement. L'existence d'une relation de traduction entre deux séquences est estimée par le cosinus des vecteurs qui leur sont associés. Il est obtenu en divisant le produit scalaire des deux vecteurs par le produit de leurs normes :

$$\sum\sum\sum \times \cdot = 22),\cos(iiiiiiyxyxyx$$

Les séquences proposées pour l'alignement sont celles qui obtiennent le cosinus le plus grand. On ne propose pas d'alignement si le meilleur cosinus est inférieur à un seuil donné (on notera que ce cosinus n'est jamais négatif puisque les composantes des vecteurs sont toujours positives).

3.3. Phase 2 : Alignement inter-document

À l'issue de première phase de résolution à l'échelle des documents, nous disposons d'un certain nombre d'hypothèses de relation de traduction entre des segments d'une langue vers une autre. Le passage à l'échelle du corpus va nous permettre de confirmer ou d'infirmer les hypothèses précédemment posées.

Pour qu'une hypothèse résiste au filtrage au niveau du corpus elle doit respecter l'un des deux critères suivant :

1. avoir été extraite d'au moins deux bi-textes ;
2. si elle n'a été trouvée que dans un seul bi-texte alors les deux séquences doivent être strictement identiques ou bien avoir une fréquence supérieure à un seuil déterminé empiriquement (6 en pratique).

Les relations de traduction sont ensuite triées en fonction du nombre d'occurrences et du nombre de documents dont elles ont été extraites.

3.4. Evaluation

Nous avons réalisé une évaluation portant sur tous les bi-textes anglais-grec. Les résultats montrent que les couples traductionnels générés sont de très bonne qualité mais encore peu nombreux. 1 066 alignements ont été proposés parmi lesquels 60 sont erronés et 32 sont incomplets. Ainsi, le taux de précision est de 91 % en considérant les résultats incomplets comme erronés. Si l'on admet des résultats incomplets comme acceptables le taux de précision est de 94 %. Pour cette évaluation, les alignements considérés comme corrects sont ceux qui présentent une équivalence traductionnelle à la fois correcte et complète, c'est-à-dire sans mots supplémentaires ou manquants, notamment dans le cas de termes ou d'expressions mis en relation.

Pour une partie des alignements proposés, on constate un recouvrement partiel entre les séquences : il s'agit de termes complexes ou d'expressions de la langue source qui ont comme équivalent un terme complexe ou une expression dans la langue cible, mais dont seule une partie a été repérée par notre algorithme. On trouve réciproquement parfois dans la langue source une partie d'un terme complexe ou d'une expression qui est mise en équivalence avec un terme complet ou une expression complète dans la langue cible. Il est intéressant de noter que les correspondances incomplètes sont parfois accompagnées de correspondances complètes (p.ex. type – διαβήτη τύπου / type 2 diabetes – διαβήτη τύπου 2, as – καθώς και / as well as – καθώς και). Il s'agit là de sous-séquences d'une séquence répétée qui ont un effectif supérieur à celui de la séquence complète et qui sont ainsi retenues (voir section 3.2.2.). Ce phénomène s'explique en partie par le fait que l'on n'adopte pas de contraintes très rigides concernant la taille des unités linguistiques à aligner. Ce choix nous offre la possibilité de prendre en compte la flexibilité de la langue et de proposer des correspondances à des niveaux différents mais il se révèle parfois problématique.

Lorsqu'un terme complexe est suivi d'une abréviation dans un des textes, c'est souvent l'abréviation qui est mise en correspondance avec le terme complexe équivalent (tout ou une partie de ce dernier) et non le terme complexe (p.ex. DM / σακχαρώδη διαβήτη). Cette situation se rencontre lorsque l'abréviation est systématiquement utilisée à la place du terme complexe dans l'un des deux textes. Dans ces cas, on a considéré les résultats comme incomplets.

Nous ne disposons pas pour le moment de mesure précise du silence, mais il est non négligeable car la méthode ne considère comme candidates à l'alignement que les séquences répétées au sein du document. Les séquences qui ne sont jamais répétées au grain document n'ont par conséquent aucune chance d'aboutir à un alignement, et cela, même si ces séquences hapax au grain document apparaissent dans plusieurs documents du corpus. Ce choix permet cependant de contrôler l'explosion combinatoire.

La nature des alignements proposés est variée. Il ne s'agit pas seulement de mots pleins mais également de mots fonctionnels (articles, conjonctions, etc.) ainsi que de termes complexes ou d'expressions. Peu de verbes sont extraits, ce qui peut entre autre s'expliquer par le fait qu'au grain document la répétition lexicale concerne davantage le contenu nominal que le contenu verbal. Malgré la liberté de la méthode quant à la longueur des séquences appariées, on constate que les alignements proposés correspondent parfois au concept de *chunk* mais que bien souvent, ils ne constituent pas des unités de nature bien identifiée. On trouve des expressions ou bien plus simplement des portions de phrases sans statut bien établi.

Parmi les couples corrects, on trouve des couples de termes relatifs à la ligne éditoriale et à la structure des sources des documents : nom de la source, sous-titres de fiche résumé, dates (par exemple Αρχ Ελλ Ιατρ – Arch Hellen Med, ΑΠΟΤΕΛΕΣΜΑΤΑ – RESULTS ΣΥΜΠΕΡΑΣΜΑΤΑ – CONCLUSIONS), des couples de termes liés au domaine médical en général : des unités de mesures, des abréviations et des termes latins (p.ex. ml – ml, SH – SH, aeruginosa – aeruginosa, vitro – vitro) et des couples de termes liés à un sous-domaine de spécialité (p.ex. κορτιζόλης – cortisol, ομοκυστεΐνης – homocysteine, τεστοστερόνης – testosterone). On note également la présence de segments de longueur variable (σε συνδυασμό με – in combination with, Αρχ Ελλ Ιατρ, 19(6), Νοέμβριος-Δεκέμβριος 2002 – Arch Hellen Med, 19(6), November-December 2002).

Enfin, le traitement d'une grande partie des variantes morphologiques en grec et en anglais améliore beaucoup les résultats. Ainsi, on a parmi les résultats des correspondances comme : (power, ισχύ~) (εξέταση~ , examination) (factor~ , κινδύνου) (protein~ , πρωτεΐνη~)(memory, μνήμη~).

Les propositions d'alignement erronées sont peu nombreuses mais on constate que dans la quasi totalité des cas, l'alignement n'est extrait que d'un ou deux documents.

4. CONCLUSION

Nous avons proposé une méthode permettant d'extraire automatiquement des relations de traduction à partir d'un corpus bilingue. Un avantage de cette méthode est qu'elle n'utilise que très peu de connaissances linguistiques et qu'elle peut donc être appliquée à des langues ne disposant que de peu de ressources électroniques disponibles ou facilement accessibles. Il s'agit d'une méthode d'analyse endogène : les connaissances ne sont extraites que du corpus. Aucun dictionnaire n'est *a priori* exigé et les affixes grammaticaux sont inférés du corpus.

La méthode est multilingue et peut s'appliquer à toute langue. Pour des raisons d'efficacité, nous avons cependant intégré des connaissances sur les frontières des mots : existence du mot graphique et liste de ponctuations.

La méthode prend en entrée un corpus de bi-textes alignés au niveau phrastique. Elle produit des alignements sous-phrastiques de longueur variable. Bien qu'il soit de plus en plus aisé de trouver des corpus bilingues alignés au niveau phrastique, et cela pour des

couples de langues très variés, il est important de noter que la qualité de l'alignement phrastique fourni en entrée n'est pas sans influence sur la qualité des résultats générés par notre méthode. Lorsque la qualité de l'alignement phrastique chute, la qualité des alignements proposés reste stable mais les propositions sont alors moins nombreuses.

La méthode que nous utilisons considère différemment le grain document et le grain corpus. Ceci permet de maîtriser l'explosion combinatoire, d'accroître la confiance dans les résultats et d'obtenir plusieurs hypothèses pour une séquence unique en cas d'homographie. Ce résultat est cependant obtenu en échange d'une augmentation du silence, silence dont nous cherchons aujourd'hui une métrique d'évaluation.

Nos travaux s'orientent vers l'alignement de segments discontinus, dans une perspective multilingue et endogène, avec un regard particulier sur la maîtrise de la combinatoire. Par ailleurs, nous souhaitons intégrer un module de résolution de coréférences qui perturbent actuellement la détection (Giguet & Lucas, 2004).

5. REFERENCES

Altenberg Bengt & Sylviane Granger. 2002. 'Recent trends in cross-linguistic lexical studies' in Altenberg & Granger (eds.), *Lexis in Conrast, Corpus-based approaches*, Amstedam / Philadelphia : John Benjamins Publishing Company, 2002, p. 3-48.

Boutsis, S., & Piperidis, S. 1998. 'Aligning clauses in parallel texts' in *Third Conference on Empirical Methods in Natural Language Processing*, 2 June, Granada, Spain, 1998, p. 17-26.

Brown P., J. Lai & R. Mercer. 1991b. 'Aligning sentences in parallel corpora' in *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, 18-21 June, Berkley, California, 1991, p. 169-176.

Gale W.A. & K.W. Church. 1991a. 'Identifying word correspondences in parallel texts' in *Fourth DARPA Speech and Natural Language Workshop*, San Mateo, California : Morgan Kaufmann, 1991, p. 152-157.

Gale W.A. & K. W. Church. 1991b. 'A Program for Aligning Sentences in Bilingual Corpora' in *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, 18-21 June, Berkley, California, 1991, p. 177-184.

Giguet E. & Lucas N. 2004. 'La détection automatique des citations et des locuteurs dans les textes informatifs' in López-Muñoz J. M., Marnette S., Rosier L. (eds.).*Le discours rapporté dans tous ses états : Question de frontières*, Paris : l'Harmattan, 2004, p. 410-418.

Harris B. 1998. 'Bi-text, a New Concept in Translation Theory', *Language Monthly*, (54), 1998, p. 8-10.

Isabelle Pierre & Susan Warwick-Armstrong. 1993. 'Les corpus bilingues : une nouvelle ressource pour le traducteur' in Bouillon, P. & Clas A. (eds.), *La Traductique : études et recherches de traduction par ordinateur*, Montréal : Les Presses de l'Université de Montréal, 1993, p. 288-306.

Kay M. & M. Röscheisen. 1993. 'Text-translation alignment', *Computational Linguistics*, March 1993, p. 121-142.

Kitamura & Matsumoto, 1996. 'Automatic extraction of word sequence correspondences in parallel corpora' in *Proc. 4th Workshop on Very Large Corpora*, Copenhagen, Denmark, 4 August 1996, p. 79-87.

Kupiec J. 1993. 'An algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora' in *Proc. of the 31st Annual Meeting of the Association of Computational Linguistics*, p. 23-30.

Papageorgiou Harris, Lambros Cranias & Stelios Piperidis. 1994. 'Automatic alignment in parallel corpora' in *Proc. 32nd Annual Meeting of the Association for Computational Linguistics*, 27-30 June, Las Cruses, New Mexico, 1994, p. 334-336.

Salkie Raphael. 2002. 'How can linguists profit from parallel corpora ?' in Lars Borin (ed.), *Parallel Corpora, Parallel Worlds : selected papers from a symposium on parallel and comparable corpora at Uppsala University*, Sweden, 22-23 April, 1999, Amsterdam, New York : Rodopi, 2002, p. 93-109.

Simard, M., G. Foster, and P. Isabelle. 1992. 'Using cognates to align sentences in bilingual corpora' in *Proc. of TMI-92*, Montréal, Québec, 1992, p. 67-81.

Smadja F. 1992. 'How to compile a bilingual collocational lexicon automatically' in *Proc. of the AAAI-92 Workshop on Statistically -based NLP Techniques*, 1992, p. 65-71.

Smadja, F., K.R. McKeown, and V. Hatzivassiloglou. 1996. 'Translating Collocations for Bilingual Lexicons : A Statistical Approach', *Computational Linguistics*, March, 1996, p. 1-38.

Tiedemann Jörg. 1993, 'Combining clues for word alignment' in *Proc. of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Budapest, Hungary, April 2003, p. 339-346.

Triantafyllou Ioannis, Iason Demiros, Christos Malavazos & Stelios Piperidis. 2000. 'An alignment architecture for Translation Memory bootstrapping' in *Proc. of the MT 2000 : Machine Translation and Multilingual Applications in the New Millenium*, Exeter, United Kingdom, 20-22 November 2000, p. 3.1-3.8.

UNE APPROCHE DE L'EXPRESSION ORALE DES ÉMOTIONS : ÉTUDE D'UN CORPUS REEL

Solange Hollard*, Mutsuko Tomokiyo, Denis Tuffelli***
***équipe GEOD, **équipe GETA**
Laboratoire CLIPS (communication langagière et interaction Personne-Système)
BP 53 38041 Grenoble Cedex 9. Grenoble
Solange.Hollard@imag.fr Mutsuko.Tomokiyo@imag.fr Denis.Tuffelli@imag.fr

1. INTRODUCTION

Nous nous sommes intéressés à l'expression des émotions dans le dialogue oral dans une double perspective. En effet, la reconnaissance et la prise en compte des émotions nous semble particulièrement intéressante pour l'interprétation pragmatique des énoncés dans les systèmes de dialogue homme-machine d'une part, et pour la représentation sémantique dans les systèmes de traduction automatique de la parole d'autre part (Tomokiyo *et al*., 2005).

Nous avons donc cherché à analyser, dans un premier temps, des énoncés réels, contenant des émotions importantes à identifier pour l'interprétation des actes de parole (Wichmann, 2002) d'une part, et facilement appréhendables par une simple audition d'autre part. Parmi ces émotions, l'impatience, voire l'irritation, nous ont semblé répondre à ces critères. Nous avons donc cherché un corpus susceptible de contenir ce type de marques. C'est alors qu'un établissement public nous a offert l'opportunité d'un accès aux messages enregistrés sur le répondeur de son service informatique. Nous avons donc sélectionné un corpus constitué d'enregistrements de ces messages. Ainsi avons-nous pu analyser l'expression orale d'usagers aux prises avec des problèmes d'informatique, comme le maniement de logiciels, l'utilisation de périphériques, les connexions au réseau.

2. LE CORPUS

Le corpus est composé d'un ensemble d'enregistrements de messages vocaux, relevés sur le répondeur du service informatique d'un établissement public français. Il s'agit uniquement de messages réels, adressés au service informatique, par des usagers membres de l'établissement. Aucun tri préalable n'a été effectué.

Ces enregistrements ont été recueillis entre octobre et décembre 2002. Ils sont regroupés dans 24 fichiers audio, d'une durée totale de 5 heures 30 (Corpus Hotline CHRU, Document CLIPS, 2003).

2.1. Contexte du recueil

L'établissement comprend plusieurs bâtiments, situés dans différents quartiers de la ville. Le service informatique est ouvert à tous les membres de l'établissement, quel que soit leur statut. Il gère tous les problèmes d'installation et d'utilisation des ordinateurs et du réseau. Il est également chargé de la gestion des cartes électroniques d'identification des

personnels. Ces cartes déterminent l'accès aux bâtiments et au réseau informatique. Les usagers qui font appel au service informatique peuvent laisser des messages sur le répondeur.

2.2. Composition du corpus

Le corpus contient deux types de messages bien distincts, qui s'adressent aux deux attributions du service informatique :

2.2.1. DEMANDES DE GESTION DE CARTES D'ACCÈS Une partie des messages concerne la gestion des cartes électroniques d'identification des personnels : demandes de validation pour de nouveaux arrivants ou pour des personnels qui rentrent après une interruption longue, demandes d'annulation en cas de départ, demandes de révision pour des cartes qui ne fonctionnent pas correctement.

2.2.2. DEMANDES D'ASSISTANCE TECHNIQUE L'autre partie des messages est constituée de demandes d'assistance technique variées. Il peut s'agir de difficultés dans le maniement d'un logiciel, dans l'utilisation des imprimantes, dans l'accès au réseau, etc.

3. ANALYSE DU CORPUS

Nous avons sélectionné les messages porteurs d'émotion. Trois auditeurs, travaillant séparément, ont écouté les messages et sélectionné ceux qui leur ont semblé porteurs d'émotion. Ceux-ci ont été transcrits, et une partie a été traitée par les logiciels Wave Edit (Akbar, 1997) et Praat (Boersma & Weenink ,1992).

Notre analyse porte sur la structure des messages, l'expression linguistique des actes de langage et les marques d'émotion.

L'analyse en actes de langage, complétée par l'analyse des stratégies discursives (Ca 04), s'effectue selon la technique mise au point au CLIPS par J. Caelen (2002), à partir du modèle de l'univers narratif de Greimas et des travaux de D. Vanderveken (1990) et D. Vernant (1992, 1997), dans un but d'interprétation pragmatique pour les systèmes de dialogue homme-machine.

3.1. Structure des messages

La structure générale des messages n'offre que peu de diversité. Elle est de la forme :

1- Phase d'ouverture : *« Bonjour… »*
2- Phase d'identification : Nom Prénom
   Service
   Poste téléphonique fixe et/ou mobile
3- Motif de l'appel : Gestion des cartes
   Problème technique
4- Localisation : logiciel, ordinateur, réseau, …
5- Exposé du problème : difficulté de fonctionnement, panne, …
6- Résumé du problème éventuellement
7- Expression de la demande : « *Pourriez-vous venir voir, pouvez-vous me rappeler, me « biper », si vous pouviez passer rapidement, …* »
8- Remerciements
9- Salutations

3.2. Marques d'émotions

Les marques d'émotion apparaissent essentiellement dans les demandes d'assistance technique. En effet, celles-ci émanent d'usagers aux prises avec des problèmes qui les gênent, voire les bloquent dans leur travail. Nous avons donc analysé les marques d'émotions qui apparaissent dans ces demandes, notamment dans les étapes 5 et 7 (exposé du problème et expression de la demande), en termes de sémantique et de phonologie.

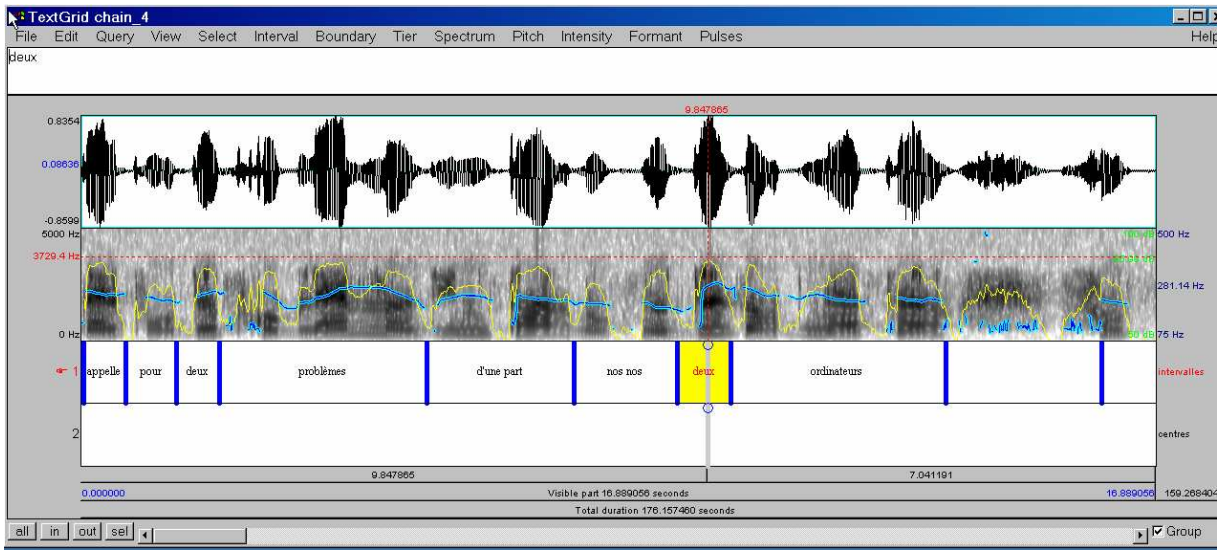
3.2.1. MARQUES SÉMANTIQUES Si la structure générale des messages est assez figée, l'expression sémantique en revanche est extrêmement variée. Une telle diversité est remarquable dans l'expression de demandes somme toute très semblables dans leur objet, émanant, qui plus est, de personnes du même milieu de travail. Cette constatation a confirmé, pour nous, l'intérêt des observations de terrain.

Les marques sémantiques peuvent être classées en quatre catégories, selon qu'elles portent sur :

- l'aspect technique du problème rencontré : « Excel est inaccessible », « les portables ne fonctionnent plus », « je n'ai plus accès à mes adresses »
- les conséquences : « je suis paralysée dans mon travail, et comme c'est un poste clef… », « on est complètement plantés, là, et on a des patients qui attendent… », « j'ai une salle d'attente pleine de monde… »
- l'état psychologique de l'émetteur : « c'est très pénible », « ici infirmière en état désespéré »,
- la demande : « Est-ce que vous pourriez nous arranger ça pour de bon ? »

3.2.2. DONNÉES PHONOLOGIQUES Dans cette première étape de notre étude, nous avons essentiellement analysé le ton (calme, neutre, élevé) et le débit (lent, normal, rapide). Nous avons effectivement relevé des élévations du ton, une accentuation de certains termes et une accélération du débit dans plusieurs énoncés. Compte tenu du contexte d'énonciation et de la spécificité du message, nous avons interprété ces modifications phonologiques comme des marques d'impatience.

La figure ci-dessous illustre la comparaison entre les deux occurrences d'un même mot, l'adjectif numéral « deux », prononcé deux fois au cours du même énoncé : « *…J'appelle pour deux problèmes, d'une part, donc nos deux ordinateurs sont euh … ne peuvent pas être démarrés suite, suite à une coupure d'électricité cette nuit* ». Le mot « deux » est prononcé la première fois de façon neutre : « *j'appelle pour <u>deux</u> problèmes…* », la seconde fois avec insistance « *…nos <u>deux</u> ordinateurs sont en panne* ». La comparaison porte sur la durée, l'énergie et le pitch, recueillis à l'aide du logiciel Praat, et de scripts spécifiques (Tuffelli, 2005).



Graphique 1 : Comparaison de durée, pitch et énergie sur un même mot prononcé de façon neutre, et avec émotion

Les mesures effectuées sur ces deux cibles sont présentées dans le tableau suivant.

Mot prononcé	Énergie moyenne (en dB)	Durée (en secondes)	Pitch moyen (en hertz)	Mesures sur le mot précédent: énergie moyenne (en dB)	Mesures sur le mot précédent: pitch moyen (en hertz)	Écart avec le mot précédent: énergie moyenne (en dB)	Écart avec le mot précédent: pitch moyen (en hertz)
Deux (neutre)	77,3322	0,6672	250,1578	70,7792	232,1596	6,5530	17,9981
DEUX (avec émotion)	81,8968	0,8428	270,6249	73,4685	208,2545	8,4284	62,3704

Tableau 1 : Mesures de durée, pitch et énergie sur un même mot prononcé de façon neutre, et avec émotion

On peut voir, dans ces illustrations, que l'énergie, le pitch et même la durée sont supérieurs dans le cas du mot prononcé avec émotion. L'écart avec l'énergie et le pitch du mot précédent est également plus élevé.

3.2.3. MARQUES SÉMANTIQUES ET MARQUES PROSODIQUES Ce corpus montre particulièrement clairement que les marques phonologiques d'émotion ne sont pas concordantes avec les marques sémantiques (Tomokiyo et al., 2005). C'est-à-dire que l'on peut observer par exemple des marques sémantiques d'impatience comme « c'est relativement énervant » ou encore « c'est très pénible » exprimées sur un ton très doux, tandis que les marques phonologiques se cantonnent sur un seul terme, moins marqué sémantiquement. Par exemple, dans l'énoncé « Il faudrait impérativement résoudre ce problème ce matin », seul le mot « matin » est accentué, tout ce qui le précède étant prononcé d'un ton neutre.

3.3. Interprétation de la non concordance sémantique/ prosodie

On peut interpréter la non concordance sémantique/prosodie en se référant aux théories ethnosociologiques (Goffman, 1974) ou cognitives.

Une première interprétation est liée à la situation des locuteurs. Il s'agit toujours d'utilisateurs face à un problème qu'ils ne peuvent résoudre, et qui demandent de l'aide au destinataire de leur message. Ils préfèrent donc s'adresser à lui aimablement, par crainte de « perdre la face » au sens de Goffman, et d'essuyer un refus.

Une autre interprétation est liée aux contraintes sociales de politesse. Il est malséant de crier ou d'utiliser des termes impolis, d'autant plus que les locuteurs s'adressent à des collègues de travail qu'ils connaissent.

Une troisième interprétation, inspirée des travaux de G. Caelen-Haumont (2004), consiste à considérer que le locuteur a le choix, pour exprimer son émotion, entre plusieurs modalités : contenu sémantique explicite ou implicite, modulation de la voix, mimiques, gestes, etc. et qu'il maîtrise ce choix. En l'occurrence, les locuteurs se contenteraient d'une seule des modalités dont ils disposent dans cette situation (sémantique, voix) pour exprimer leur émotion.

4. CONCLUSION

Le travail d'analyse de ce corpus réel doit être approfondi, notamment dans deux directions : l'optimisation de la représentation sémantique dans le projet UNL (Universal networking langage) dont l'équipe Geta du laboratoire est partenaire, et la recherche axée sur l'interprétation pragmatique en dialogue homme-machine (Xuereb & Caelen, 2004a et b), notamment sur les relations rhétoriques (Xuereb & Caelen 2005) dans les actes de langage, en cours dans l'équipe GEOD.

5. REFERENCES

Akbar, M. 1997. Wave Edit, logiciel d'édition et d'étiquetage, équipe GEOD (Groupe d'étude de l'oral et du dialogue), Laboratoire CLIPS (Communication langagière et interaction personne-Système), BP 53 38041 Grenoble Cedex 9.

Boersma, P. et Weenink, D., 1992-2005. *PRAAT, Doing Phonetics by Computer*, www.Praat.org

Caelen J. 2002. 'Modèles formels de dialogue', in Le Maître J. (ed.) *Actes des 2èmes assises du GdR I3, Information, Interaction Intelligence*, Cépadues Editions, p. 31-58.

Caelen J. (à paraître). 'Dialogue strategies', in Caelen, J., Vanderveken D. Vernant D. (eds.), *Logic and Dialog*, Netherlands, Dordrecht, Kluwer.

Caelen J. 1997. *Interaction verbale*, Cépadues Editions.

Caelen-Haumont G. (à paraître). 'Les états émotionnels et la Prosodie : paradigmes, modèles, paramètres.' in ouvrage collectif, *Phonologie et phonétique : Approches contemporaines*. Paris : Hermès.

Goffman E. *Les rites d'interaction*, Paris : Éditions de Minuit, 1974.

Tomokiyo M., Chollet, G. Hollard S. 2005. 'Studies of Emotional Expressions' in Cerdeñosa J., Gelbukh A., Tovar E. (eds.), *Oral Dialogues, Research on computer science, Universal Networking Language, Advances in theory and application*, Mexico, p. 286-299.

Tuffelli, D. 2005., Scripts Praat de segmentation et d'étiquetage d'un signal de parole : Ini : Ecoutes et segmentations initiales ; Rep : reprise d'une segmentation effectuée auparavant, équipe GEOD (Groupe d'étude de l'oral et du dialogue), Laboratoire CLIPS (Communication langagière et interaction personne-Système), BP 53 38041 Grenoble Cedex 9.

Vanderveken, D. 1990. *Meaning and speech acts*, Cambridge University Press.

Vernant D. 1992. ‘Approche actionnelle et modèle projectif du dialogue informatif.’ in *Du dialogue, recherches sur la philosophie du langage*, Paris : Vrin, n°14, p. 265-314.

Vernant D. 1997. Du discours à l’action : études pragmatiques. Paris : PUF.

Wichmann A. 2002 ‘Attitudinal Intonation and the Inferential Process’, International Conference Speech Prosody, www.lpl.univ-aix.fr/projects/aix02/sp2002/pdf/wichmann.pdf

Xuereb A. and Caelen J. 2004. ‘Un modèle d'interprétation pragmatique en dialogue homme-machine basé sur la SDRT’, in *XIème Conférence sur le Traitement Automatique du Langage Naturel (TALN'04)*, Fès, Maroc, 19-21 avril 2004, p. 505-514.

Xuereb A. and Caelen J. 2004. ‘L'interprétation pragmatique en dialogue homme-machine finalisé’, in *Journées JSM'04*, Lyon, mars 2004

Xuereb A. and Caelen J. 2005. ‘Speech Acts and Rhetorical Relations in Man-Machine Dialogue’, in *Séminaire Dialogue et Logique*, Nancy : Presses Universitaires de Nancy.

EXTRACTION AUTOMATIQUE DE TERMES A PARTIR DE TEXTES POLONAIS

Anna Kupść
CNRS, Loria, UMR 7503, Nancy et
Académie Polonaise des Sciences, Institut d'Informatique, Varsovie

1. INTRODUCTION

Cet article présente nos études sur l'extraction (acquisition) automatique de termes à partir de textes spécialisés écrits en polonais. Notre travail a pour objectif le développement de ressources linguistiques. Les outils pour l'aquisition de termes sont assez bien décrits pour l'anglais ou le français (Smadja, 1993 ; Daille 1996 ; Justeson & Katz, 1991 ; Bourigault, 1994 ; Bowden *et al*., 1998), mais la question reste peu étudiée pour les langues slaves, qui se caractérisent par leur flexion plus riche et leur ordre des mots plus libre. Pour le polonais, nous avons adopté la méthode de Frantzi *et al*. (2002), déjà utilisée pour deux autres langues slaves : le slovène (Vintar, 2004) et le serbe (Nenadic *et al*., 2003). La méthode permet de découvrir des termes complexes (comprenant plusieurs mots) et enchâssés, en utilisant une approche hybride, appliquée au corpus specialisé. L'article est organisé de la manière suivante : la section 2 présente le corpus, la méthode appliquée est décrite dans la section 3, les résultats d'extraction et les problèmes rencontrés sont examinés dans les sections 4 et 5, respectivement ; la conclusion dans la section 6 termine l'article.

2. CORPUS

Le corpus que nous avons utilisé comprend des dossiers de patients diabétiques d'un hôpital polonais, rassemblés pendant une période de quatre ans. Les données des trois premières années (2001-2003) ont servi de corpus d'entraînement et les dossiers de la dernière année (2004) étaient utilisés pour l'évaluation finale (corpus de test). Les textes originaux sont écrits en Word et, pour le traitement, ils ont été convertis en texte clair, encodé en UTF-8. Les dossiers contiennent quelques fautes d'orthographe mais leur qualité est en général bonne et aucun pré-traitement n'a été appliqué. Le corpus contient des textes bruts : 46 449 mots dans le corpus d'entraînement et 13 596 dans le corpus de test.

3. METHODE

Nous avons adopté l'approche *C/NC-value* proposé par Frantzi *et al*. (2002) qui combine des méthodes linguistiques et statistiques : d'abord, des patrons syntaxiques sont utilisés pour identifier des termes candidats (syntagmes nominaux) dans les textes annotés ; ensuite, les candidats retenus sont classés selon leur stabilité par rapport aux termes les plus longs (*C-value*) et aussi selon leur importance par rapport au contexte dans lequel ils apparaissent (*NC-value*). Cette technique a été appliquée avec succès à l'anglais, au japonais, et à deux langues slaves, le slovène et le serbe ; ici nous l'utilisons pour le polonais, une autre langue slave.

Pour le polonais, nous ne disposons pas de textes annotés. Alors, nous avons utilisé un analyseur de surface sur les textes bruts. La description précise des tâches effectuées pour le polonais est présentée dans les sections qui suivent.

3.1. Partie linguistique

Pour l'identification des termes candidats, nous avons utilisé SProUT (Shallow Processing with Unification and Typed Feature Structures), une plateforme d'analyse de surface, complétée d'un analyseur morphologique du polonais (Piskorski *et al*., 2004). SProUT combine des techniques à états finis avec un formalisme d'unification. Ceci permet de construire des syntagmes basés sur mots-clés et indices syntaxiques. Nous avons développé une grammaire de surface pour les syntagmes nominaux du polonais. La grammaire comprend 8 règles qui tiennent compte des principales structures nominales : l'accord en nombre, en genre et en cas entre le nom et l'adjectif (1), et les syntagmes à complément génitif (2), où uniquement le nom-tête *poziom* « niveau » peut changer de cas tandis que la forme du complément est invariable.

1) *diecie* *cukrzycowej*
   régime$_{fem.sg.dat}$ diabétique$_{fem.sg.dat}$
   « au régime diabétique »
2) *poziom* *glukozy*
   niveau$_{msc.sg.nom}$ glucose$_{fem.sg.gen}$
   « niveau de glucose »

Les règles linguistiques reconnaissent beaucoup de syntagmes qui ne sont pas des termes. Par exemple, *dotychczasowe leczenie* « traitement précédent » est reconnu comme un candidat mais ce n'est pas un terme. Pour réduire le nombre de candidats, nous n'avons appliqué les règles qu'aux syntagmes comprenant des mots « 'spécialisés » (cf. Drouin, 2003). Nous avons créé une liste de mots rencontrés plus fréquemment dans le corpus médical que dans les textes généraux. Comme corpus de référence, nous avons utilisé trois années du quotidien *Rzeczpospolita* et nous avons calculé la fréquence relative (Damerau, 1993) par rapport aux textes médicaux. Ensuite, cette liste a été filtrée pour éliminer les verbes, les adverbes, les mots fonctionnels, les chiffres, les symboles, les signes de ponctuation, etc. Par conséquent, les règles ne sont appliquées qu'aux syntagmes contenant les noms et les adjectifs spécifiques au domaine.

3.2. Post-traitement

SProUT ne comporte pas de tagger, ni de déducteur (guesser) et les mots absents du lexique de l'analyseur morphologique ne sont pas reconnus. En fait, beaucoup de mots inconnus font partie des termes, par exemple les mots composés ou les noms latins des maladies et des médicaments. Alors, nous tenons compte aussi des séquences de mots qui ne sont pas reconnues par les règles décrites dans la section 3.1. Parce qu'on ne peut pas détecter les mots inconnus directement, nous faisons un filtrage négatif : nous appliquons des « règles stop » et nous rejetons les mots fonctionnels, les adjectifs non-attachés, les adverbes, les verbes, les chiffres, les symboles, les signes de ponctuation, etc. Ensuite, les mots inconnus sont incorporés aux syntagmes nominaux extraits. Pour le groupement, nous utilisons deux heuristiques :

1. s'il y a un SN adjacent au groupe de mots inconnus, ces mots sont ajoutés à ce syntagme ;

2. autrement, le groupe de mots inconnus est considéré comme un nouveau terme.

Nous notons aussi les contextes dans lesquels les termes apparaissent : selon Sager (1978), la modification des termes est plus restreinte que pour les non-termes. Nous reconnaissons trois types de modificateurs / contextes : les noms, les adjectifs et les verbes.

Le résultat de cette tâche est une liste de fréquence des termes candidats avec leurs contextes.

3.3. Partie statistique

La méthode *C/NC-value* de Frantzi *et al*. (2000) permet d'identifier des termes complexes (comprenant plusieurs mots) mais aussi des termes enchâssés. Par exemple, elle peut indiquer que la séquence $biochemiczne_A$ $zapalenie_N$ $trzustki_N$ ne comprend que deux termes : $biochemiczne_A$ $zapalenie_N$ $trzustki_N$ « pancréatie biochimique » et $zapalenie_N$ $trzustki_N$ « pancréatie », mais que $biochemiczne_A$ $zapalenie_N$ « inflammation biochimique » n'est pas un terme, bien que ce soit un syntagme nominal.

La mesure *C-value* vérifie la stabilité des termes par rapport aux termes les plus longs et *NC-value* les classe selon leur importance par rapport au contexte.

3.3.1. C-VALUE La mesure est basée sur la fréquence du terme t dans le corpus [f(t)], la fréquence du terme comme un terme enchâssé [nt], le nombre de termes enchaînant le terme [Nt] et la longueur du terme (nombre de mots) [l] :

$$C\text{-}value(t) = \begin{cases} \log_2 l \, f(t) & \text{si } t \text{ n'est pas enchâssé} \\ \log_2 l \, (f(t) - n_t / N_t) & \text{sinon} \end{cases}$$

3.3.2. NC-VALUE Cette mesure reprend l'observation de Sager (1978) que la modification des termes est plus restreinte que celle des non-termes. La modification est entendue au sens large comme un contexte entourant un terme. Selon Greffenstette (1994), les noms, les verbes et les adjectifs sont les environnements où les termes apparaissent le plus souvent. Ces trois contextes ont été notés dans la phase décrite dans la section 3.2.

Frantzi *et al*. (2000) associent un poids avec chaque mot-contexte [*w*] :

$$weight(w) = Tw / n$$

T_w est le nombre de termes dans le contexte *w*, et *n* est le nombre total de termes. Le poids du contexte est ensuite utilisé pour re-classer les termes découverts par la *C-value* par rapport à leur importance : les termes plus représentatifs se trouvent plus souvent dans des contextes plus importants. La formule pour calculer la *NC-value* est présentée ci-dessous :

$$NC\text{-}value(t) = 0.8\, C\text{-}value(t) + 0.2 \sum_{w \in Ct} ft(w)\, weight(w)$$

C_t – l'ensemble des différents contextes du terme *t*, $f_t(w)$ – la fréquence de *w* comme un contexte de *t*. La pondération de la *C-value* et du poids de contexte, 0.8 et 0.2, a été établie expérimentalement (cf. Frantzi *et al*., 2000).

4. EXPERIMENTATION

Nous avons évalué le système sur les corpus d'entraînement et de test. Nous n'avons pas utilisé de seuil d'occurrence à cause de la petite taille des corpus. Les termes extraits

comprennent des termes simples et complexes. L'évaluation a été faite de façon manuelle ; nous avons utilisé des valeurs « relatives » de précision et de rappel :

$$\text{précision} = \frac{\text{termes corrects}}{\text{touts les termes extraits}}$$

$$\text{rappel} = \frac{\text{fréquance de termes corrects}}{\text{fréquance de touts les termes extraits}}$$

Nous avons mesuré la performance du système sur la liste complète des termes. Les résultats généraux sont présentés dans le tableau 1. Nous avons aussi comparé les résultats obtenus pour les 100 premiers éléments de la liste, classés par la *C/NC-value* et selon la fréquence. La comparaison des résultats pour les 100 termes initiaux est présentée dans le tableau 2.

	entraînement	test
termes extraits : total	3205	1293
C/NC-value	3095	1251
précision	52.3 %	53.8 %
rappel	60.6 %	59.5 %

Tableau 1 : Résultats généraux

	C/NC-value		Fréquence	
	entraînement	Test	entraînement	test
précision	68 %	67 %	66 %	65 %
rappel	61.75 %	59.44 %	60.63 %	60.87 %

Tableau 2 : Résultats pour les 100 premiers éléments

Les résultats généraux sont comparables à ceux obtenus pour le slovène : Vintar (2004) rapporte une précision de 51 % et un rappel de 65 %. Le tableau 2 montre que le classement des termes initiaux est effectivement meilleur que celui pour toute la liste : au début de la liste, environ 16 % d'éléments de plus (précision) sont des vrais termes. Ce résultat est aussi légèrement supérieur au classement par la fréquence ; on reviendra à la discussion de la performance dans la section 5.

La liste des 25 premiers termes reconnus par le système est présentée dans le tableau 3.

rang	correct / incorrect	terme	traduction
1	1	insulina	insuline
2	0	mg	mg
3	0	norma	norme
4	0	ml	ml
5	0	leczenie	traitement
6	1	dobowy profil glikemii	profil journalier de glycémie
7	1	humalog	humalog
8	0	mmol	mmol
9	1	ekg	électrocardiogramme
10	1	posiew mocz	analyse d'urines
11	1	dieta cukrzycowy	régime diabétique
12	1	rr	tension artérielle
13	0	daleki leczenie	traitement suivant
14	0	stan	état
15	0	kg	kg
16	1	leukocyty	leucocytes
17	0	wg	selon
18	1	humulin	humulin
19	1	erytrocyty	érythrocytes
20	1	insulatard	insulatard
21	1	publiczny zakład opieka zdrowotny	établissement de soins médicaux publiques
22	1	wydział lekarski akademia medyczny	département médical de l'académie de médecine
23	0	dl	dl
24	0	glikemia mg	glycémie mg
25	1	pacjentka	patiente

Tableau 3 : 25 premiers termes

5. PROBLÈMES

Bien que les résultats présentés dans la section précédente soient raisonnables, la performance n'est pas optimale et on pourrait faire mieux. Nous avons identifié deux problèmes principaux liés au traitement automatique des textes en langue naturelle et à la normalisation des termes.

5.1. Problèmes de TAL

Les heuristiques pour traiter les mots inconnus peuvent en fait détruire les syntagmes reconnus par les règles grammaticaux.

(3) *w [tylnym biegunie] [mikroaneuryzmaty*]

'dans le pôle derrière microaneurismates'

Le SN *tylnym biegunie* « le pôle derrière » en (3) est identifié comme un syntagme nominal par le système, alors que le mot *mikroaneuryzmaty* « microaneurismates » est

inconnu. Parce que les deux éléments sont adjacents, les heuristiques (section 3.2) les groupent ensemble, bien qu'ils ne forment ni un syntagme ni un terme.

Le groupement des mots inconnus ne donne pas toujours un nouveau terme non plus. Cette situation est illustrée en (4) :

(4) *[glikemia] [mm]*

'glycèmie mm'

Ni « glycèmie » ni « mm » n'est reconnu par l'analyseur morphologique. Comme dans l'exemple précédent, les deux mots sont adjacents et ils sont donc considérés comme un nouveau terme par les heuristiques.

Un autre type de problème est posé par l'ambiguïté des mots reconnus. Par exemple, *bez* appartient à deux catégories grammaticales : nom (« lilas » ou bien la forme génitive plurielle du nom *beza* « macaron ») et préposition (« sans »). À cause de son interprétation nominale, il fait souvent partie des syntagmes nominaux.

Ces difficultés pourraient être évitées si l'analyseur morphologique était remplacé par un tagger ou par un guesser.

5.2. Normalisation

La flexion en polonais s'applique également aux formes complexes. L'exemple (5) présente quelques formes fléchies en cas du terme « régime diabétique » :

(5) *dieta cukrzycowa* (nominatif)

diety cukrzycowej (génitif)

diecie cukrzycowej (datif, locatif)

dietę cukrzycową (accusatif)

Chaque mot dans un terme peut être fléchi, soit en accord avec le mot-tête, (1) et (5), soit indépendamment, (2). La variation des composants pose problème pour le calcul de fréquence basé sur la forme de surface : le terme en (5) a plusieurs formes et, si elles sont comptées séparément, cela donne des résultats faussés.

Nous avons adopté une solution temporaire qui consiste à utiliser la forme de base de chaque mot (voir les résultats du tableau 3). Toutes les variantes en (5) sont alors représentées par la séquence *dieta cukrzycowy*. Malheureusement, cette solution n'est pas applicable aux termes qui contiennent des mots inconnus : ils ne sont pas reconnus par l'analyseur morphologique, donc leurs formes restent distinctes.

D'autre part, dans beaucoup de cas la séquence des formes de base ne correspond pas à la forme de base du terme. Par example, *dieta cukrzycowy* ne donne pas le nominatif en (5), parce que la forme de base des adjectifs est masculine (*cukrzycowy*) et on perd l'accord en genre avec le nom *dieta* « régime » qui est féminin en polonais.

Pour résoudre ces problèmes il faudra adopter un modèle flexionnel des formes complexes, ou bien utiliser des méthodes statistiques de groupement (angl. *clustering*).

6. CONCLUSION

Nous avons présenté un système hybride pour l'acquisition automatique des termes à partir de textes en polonais. Nous avons adopté une méthode utilisée précédemment pour d'autres langues slaves. Les résultats obtenus pour le polonais sont comparables à ceux du slovène, ce qui confirme la portabilité de la méthode.

La performance du système est raisonnable mais elle pourrait être améliorée. Les plus importants problèmes touchent au traitement des mots inconnus et de l'ambiguïté, mais aussi à la normalisation morphologique des termes, essentielle pour le calcul de fréquence. Nous allons étudier ces problèmes dans les versions ultérieures du système.

7. REMERCIEMENTS

Le travail présenté dans cet article a été en partie subventionné par le projet national du Comité de Recherche Scientifique (KBN) n° 3 T11C 007 27.

8. REFERENCES

Bourigault, D. 1994. LEXTER un Logiciel d'Extraction de Terminologie. Application à l'extraction des connaissances à partir de textes. Thèse de doctorat, École des Hautes Études en Sciences Sociales, Paris.

Bowden, P., Lindsay, E., et Halstead, P. 1998. 'Automatic acronym acquisition in a knowledge extraction program' in Proceedings of COMPUTERM, the First Workshop on Computational Terminology, Montreal.

Daille, B. 1996. 'ACABIT : une maquette d'aide à la construction automatique de banques terminologiques' in Clas A., Thoiron P. et Béjoit (eds.), Lexicomatique et Dictionnairitique, FMA : Beyrouth, p. 123-136.

Damerau, F. J. 1993. 'Generating and evaluating domain-oriented multi-word terms from texts', Information Processing and Management, 29, p. 433-447.

Drouin, P. 2003. 'Term extraction using non-technical corpora as a point of leverage', Terminology, 9(1), p. 99-117.

Frantzi K., Ananiadou S., Mima H. 2000. 'Automatic recognition of multi-word terms : the C-value/NC-value method', International Journal of Digital Libraries.

Greffenstette, G. 1994. Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers.

Justeson, J. S. et Katz, S. M. 1991. 'Co-occurrences of antonymous adjectives and their context', Computational Linguistics, 17, p. 1-19.

Nenadic G., Spasic I, et Ananiadou S. 2003. 'Morpho-Syntactic Clues for Terminological Processing in Serbian' in Proceedings of EACL Workshop on Morphological Processing of Slavic Languages, Budapest, Hungary.

Piskorski, J., Homola, P., Marciniak, M., Mykowiecka, A., Przepiórkowski, A., Woliński, M. 2004. *Information Extraction for Polish Using the SProUT Platform. The proceedings of Intelligent Information Systems 2004 (New Trends in Intelligent Information Processing and Web Mining)*, Zakopane.

Sager J. C. 1978. 'Commentary by Prof. Juan Carlos Sager' in Actes Table Ronde sur les Problemes du Decoupage du Termes, Montreal, Quebec, 1978.

Smadja, F. 1993. 'XTRACT : An overview', Computers and the Humanities, 26, p. 399-413.

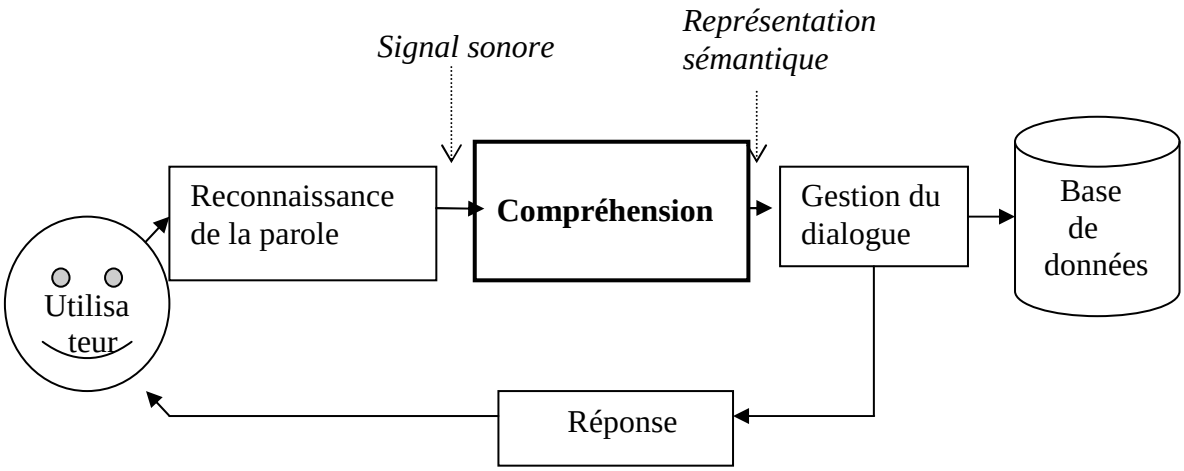
Vintar S. 2004. 'Comparative Evaluation of C-Value in the Treatment of Nested Terms' in Proceedings of the Methodologies and Evaluation of Multiword Units in Real-word Applications, LREC.

CORPUS DE DIALOGUE HOMME/MACHINE : ANNOTATION SÉMANTIQUE ET COMPRÉHENSION

Jeanne Villaneau[1] et Sylvain Lamprier[2]
(1) VALORIA - UBS, (2) M2R IHM à l'UBS

1. INTRODUCTION

Grâce aux progrès récents de la reconnaissance vocale, les systèmes de Dialogues Oraux Homme/Machine (DOHM) deviennent progressivement accessibles au grand public. La plupart de ces systèmes permettent d'interroger une base de données et leur architecture générale correspond au schéma suivant :



Malgré des avancées significatives, les systèmes de DOHM ne sont actuellement destinés qu'à des tâches extrêmement précises : interrogation sur les horaires d'avion ou de train, informations météo, etc. Par ailleurs, le dialogue y est généralement dirigé par la machine et peu d'initiatives sont laissées à leurs utilisateurs qui doivent souvent se contenter de répondre aux questions qui leur sont posées. Pour réaliser des systèmes permettant des interrogations dans des domaines un peu plus larges ou un dialogue plus naturel, des progrès importants sont nécessaires. Améliorer la compréhension automatique est l'une de ces nécessités.

Après avoir donné un aperçu rapide des problèmes posés par la compréhension automatique et de ce que la linguistique de corpus peut apporter dans ce domaine, nous décrivons dans cet article un projet ministériel d'évaluation de la compréhension, pour lequel a été recueilli et annoté un corpus de dialogue oral ; nous essayons ensuite de circonscrire et préciser les questions soulevées par cette expérience.

2. COMPRÉHENSION AUTOMATIQUE ET LINGUISTIQUE DE CORPUS

La compréhension automatique des énoncés obtenus en sortie de la reconnaissance vocale est un élément central des systèmes de DOHM. Selon l'architecture présentée précédemment, la tâche du module de compréhension est de *transformer une liste de mots en une représentation sémantique utilisable* par le gestionnaire du dialogue. Même ainsi précisée, cette tâche peut être conçue de différentes manières : elle dépend largement de la conception générale du système et, en particulier, de la façon dont est envisagée la gestion du dialogue. Dans tous les cas, elle est considérée comme extrêmement complexe : le fort taux d'erreurs engendrées par la reconnaissance vocale - dans des conditions « normales » d'utilisation, on peut s'attendre à des taux d'erreurs sur les mots allant jusqu'à 30 % - et les phénomènes spécifiques à l'expression orale spontanée (« réparations », souplesse de la syntaxe) rendent impossible une analyse syntaxique complète des énoncés analysés.
Après avoir donné deux exemples pour illustrer la diversité des approches utilisées, nous montrerons le rôle de la linguistique de corpus dans le domaine.

2.1 Exemples d'approches

Les formes mêmes des représentations sémantiques rendues dépendent fortement des systèmes ; il en est de même de l'analyse des énoncés qui produisent ces formes. Aucune des différentes approches utilisées n'a encore clairement démontré sa supériorité. Deux d'entre elles sont présentées ci-après pour illustrer leur diversité.

2.1.1 APPROCHES SÉLECTIVES Ces familles d'approches sont en général liées à la définition préalable de schémas sémantiques susceptibles de représenter toutes les requêtes possibles. La compréhension consiste alors à détecter des mots ou suites de mots-clefs (segments sémantiques) pour déterminer la nature exacte de la requête et instancier ses différents paramètres.

Par exemple, dans le système MASK destiné à renseigner sur les horaires de train (Lamel, 1995), l'analyse de l'énoncé

« je veux aller demain matin de Paris à Marseille en passant par Lyon »

repose entièrement sur la détection des segments soulignés. Par exemple, l'expression *« de Paris »* est interprétée sans ambiguïté comme une détermination de la ville de départ, de par la présence de la préposition *« de »*. Ainsi, la représentation du sens de l'énoncé correspond au schéma sémantique suivant :

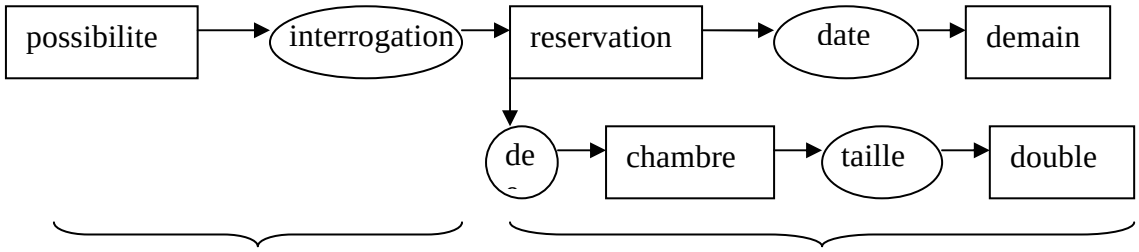
<train-heure>	ville-départ	Paris
	ville-arrivée	Marseille
	ville-escale	Lyon
	jour-relatif	demain
	période-journée	matin

Ces méthodes sont suffisantes et efficaces pour des domaines très étroits, où l'ambiguïté sémantique est réduite. Elles deviennent difficiles à appliquer lorsque le domaine d'application s'élargit (Allen, 2001).

2.1.2 APPROCHE LOGIQUE Cette approche a pour objectif de permettre de représenter le sens dans des domaines plus larges, sans avoir recours à des schémas sémantiques prédéfinis. Elle repose sur le principe qu'il est possible de définir une ontologie du domaine dans laquelle sont définis les liens sémantiques possibles entre les différents

concepts de l'application. La représentation sémantique rendue est un graphe conceptuel (Sowa, 2000) qui explicite les liens (relations conceptuelles) entre les différents concepts de l'application identifiés dans l'énoncé ; elle est inspirée de la théorie des actes de langage (Austin, 1962) et de la logique illocutoire (Vanderveken, 2001). L'analyse s'appuie conjointement sur des critères syntaxiques et sémantico-pragmatiques et elle se fait en plusieurs étapes, d'une manière ascendante et incrémentale (Aït-Mokhtar, 2002) : l'association des concepts identifiés dans l'énoncé permet de lever progressivement les ambiguïtés pour aboutir à la formule finale. Le système LOGUS (Villaneau, 2004) repose sur ces principes et le domaine test choisi est le renseignement touristique.

Par exemple, à partir de l'énoncé *« est-ce que je peux réserver une chambre euh une chambre double pour demain »*, LOGUS construit une formule logique qui correspond au graphe conceptuel suivant :



Acte de langage (indication de)

Chaîne d'objets

2.2 Compréhension automatique et corpus

Pour permettre le développement de la compréhension automatique dans les systèmes de DOHM, les corpus oraux sont indispensables à double titre.

- Ils sont indispensables pour la réalisation même des systèmes. Savoir comment leurs utilisateurs vont s'exprimer est difficile à prévoir : il n'est pas si courant de s'adresser à une machine. En général, pour la conception d'un nouveau système, connaître le mode d'expression de ses utilisateurs s'apparente au problème de la poule et de l'œuf (Glass, 1999) : il s'agit en effet de déterminer les réactions d'utilisateurs potentiels d'un système avant même qu'il n'existe… Comme on ne s'exprime pas de la même manière lorsqu'on s'adresse à un être humain que lorsqu'on s'adresse à une machine, il est essentiel, pour avoir des corpus représentatifs, de recueillir des échantillons de dialogue Homme/Machine. Une méthode classique consiste à recueillir un premier corpus en utilisant la technique dite du « magicien d'Oz ». Celle-ci consiste à faire croire à des cobayes qu'ils utilisent un véritable système alors que le rôle de la machine est joué en réalité par des humains qui simulent les réactions de cette machine. À partir de ce premier corpus, un prototype du système est réalisé, qui permet de recueillir de vrais corpus, qui eux-mêmes permettent d'apporter des améliorations au système, etc. selon une méthode incrémentale. Ce procédé risque tout de même d'introduire un biais expérimental : les utilisateurs s'adaptant très rapidement aux défauts d'un système, il n'est pas certain que les énoncés posant problème apparaissent dans le corpus (l'orateur risque de contourner les obstacles). Les corpus recueillis en situation ne

sont donc pas représentatifs en soi d'un mode d'expression idéal Homme/Machine mais d'un mode d'expression lié à un système particulier et à ses défauts.

- Les corpus de dialogue sont également indispensables pour une tâche complexe mais néanmoins indispensable : l'évaluation des systèmes. Comme il a été dit précédemment, améliorer la compréhension automatique est un enjeu crucial pour le développement des systèmes de DOHM et, actuellement, on ignore quelles sont, parmi les approches utilisées, celles qui sont les plus efficaces. Or, évaluer la pertinence et l'efficacité d'un système de compréhension est un problème extrêmement complexe (Devillers, 2002) pour lequel un certain nombre de paradigmes ont été proposés (Minker, 1998 ; Antoine, 2002). La première difficulté à surmonter est de disposer de corpus de dialogue de tests représentatifs de la tâche et du domaine.

3. LE PROJET MEDIA

Le consortium MEDIA-EVALDA du projet ministériel français TECHNOLANGUE s'est donné pour objectif de définir et de tester une méthodologie pour l'évaluation des systèmes de compréhension destinés au Dialogue Oral Homme/Machine ; le groupe ELDA a été chargé de coordonner les différentes campagnes du projet EVALDA dont MEDIA est un sous-projet. Il regroupe plusieurs (une grande partie des) labos français qui travaillent dans le domaine.

3.1. Les étapes du projet

Le consortium a tout d'abord fait le choix d'un domaine d'application : la réservation hôtelière ; il a ensuite recueilli un corpus de dialogue Homme-Machine représentatif de cette tâche. Ce corpus a fait l'objet d'une annotation sémantique manuelle de référence pour une campagne d'évaluation hors-contexte de dialogue (Bonneau-Maynard, 2005). Une seconde campagne d'évaluation en contexte de dialogue est envisagée, si le temps imparti au projet MEDIA le permet.

3.2. Le corpus

3.2.1. RECUEIL DU CORPUS Les membres du consortium ont proposé un certain nombre de scénarii de réservation hôtelière. Par exemple : « *vous cherchez un hôtel le plus près possible de l'aéroport toulousain pour que vous puissiez partir très tôt le lendemain matin. Renseignez-vous sur les attractions touristiques proches de l'hôtel pour occuper votre après-midi (pas de préférence particulière mais vous voulez pouvoir faire quelque chose)* ». La technique du magicien d'Oz a été utilisée : les « cobayes » croyaient qu'ils s'adressaient à une machine. Une série de consignes très précises avait été distribuée aux compères qui jouaient le rôle de la machine, afin de rendre leurs réactions systématiques et crédibles.

3.2.2. CONTENU DU CORPUS 250 utilisateurs du système ont suivi chacun 5 scenarii de réservation différents. Le corpus est composé de 1 257 dialogues, ce qui correspond à environ 70 heures d'enregistrements. Ces dialogues ont ensuite fait l'objet d'une retranscription manuelle par ELDA. La complexité des énoncés utilisateurs est très variable comme le montrent les exemples donnés ci-après.

Les énoncés utilisateurs peuvent être très simples comme dans la suite d'échanges ci-dessous où ils ne correspondent qu'à de brèves réponses :

- *bienvenue sur le serveur MEDIA système d'informations touristiques et de réservation d'hôtel souhaitez-vous écouter le guide ?*
- **non**
- *quelle information désirez-vous ?*
- **réserver une chambre d'hôtel**
- *dans quelle ville désirez-vous réserver une chambre d'hôtel ?*
- **Grenoble Marseille et Albi**
- *à Grenoble à quelle période souhaitez-vous y séjourner ?*
- **du trente juin au trois juillet**
- *combien de chambres souhaitez-vous réserver à Grenoble ?*
- **une chambre simple avec un parking privé**
- *merci de patienter je procède à votre recherche à Grenoble* **…**
- etc.

Certains des énoncés utilisateurs peuvent néanmoins être qualifiés de complexes, au sens où ils peuvent poser problème à des systèmes de compréhension automatique et ce, pour des raisons diverses, comme le montrent les exemples suivants :

- **je voudrais faire une réservation du cinq au huit mars trois nuits à Châteauroux ; du huit mars au neuf mars à Amiens avec dix-huit chambres individuelles quatre-vingts euros maximum et je voudrais des douches dans les chambres je voudrais que l'hôtel soit situé près d'un parc**
- **réserver une chambre simple du premier au six décembre dans trois endroits différents**
- **alors il me faudrait encore deux chambres doubles et une chambre simple mais cette fois-ci du vingt-cinq au vingt-sept mai à Épernay**
- **chat est-ce que ça pose un problème**

3.2.3. REPRÉSENTATION SÉMANTIQUE La forme même de la représentation sémantique de référence a été l'un des grands problèmes qu'a dû gérer le consortium. Pour des raisons diverses, où la faisabilité a joué une large part, le choix s'est porté sur une annotation sémantique des segments « porteurs de sens » au sens de l'application. La forme de l'annotation correspond à un schéma fréquemment utilisé dans les méthodes sélectives : chaque segment correspond à un triplet <mode, attribut, valeur> où :

mode : +, -, ?, ~ indiquent si le segment correspond à une affirmation, un refus, une interrogation ou une possibilité.

Attribut : donne la nature du concept

valeur : précise la valeur du concept.

L'annotation suit les segments dans l'ordre où ils apparaissent dans l'énoncé. Dans l'exemple ci-dessous, *nombre-chambre, chambre-type, connectprop* sont des attributs, *2, double,* … sont des valeurs ; + est le mode par défaut.

« alors il me faudrait deux chambres doubles et une simple »

deux chambres : +:nombre-chambre:2

doubles	:	+:chambre-type:double
et	:	+:connectprop:addition
une	:	+:nombre-chambre:1
simple	:	+:chambre-type:simple

Les trois exemples donnés ci-après illustrent le problème du mode et des références. Le premier donne un exemple de mode interrogatif. Le deuxième illustre comment sont indiquées les références (destinées à être résolues dans l'annotation en contexte). Il montre par ailleurs la difficulté de distinguer dans la transcription écrite le mode interrogatif du mode affirmatif ; pour l'annotation manuelle, la désambiguïsation s'est faite par écoute de la prosodie. Le troisième illustre l'une des difficultés auxquelles est confrontée ce type d'annotation : le traitement des réparations. La double annotation de l'expression « *le tarif les tarifs* » montre que la réparation n'est pas « traitée » par l'annotation.

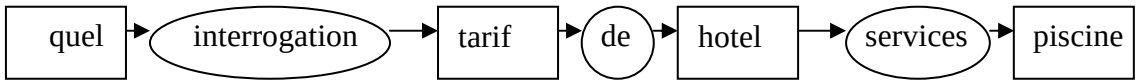
1. « euh oui est-ce que je peux ramener mon chien tout d' abord »
   REF: +:reponse:oui
   ?:hotel-services:animaux
2. *« le prix »*
   REF: ?:lienref-coref:singulier
   REF : ?:objet:paiement-montant
3. « euh quel est le tarif les tarifs pour ces deux hôtels »
   REF : ?:lienref-coref:singulier
   ?:objet:paiement-montant-chambre
   ?:lienref-coref:pluriel
   ?:objet:paiement-montant-chambre
   +:lienref-coref:pluriel
   +:nombre-hotel:2

Pour parvenir à des annotations normalisées, un manuel d'annotation a été élaboré. Le taux d'accord entre annotateurs a progressivement atteint 85 %, ce qui est normalement considéré comme un bon score pour ce type de tâches.

3.3. La campagne d'évaluation hors-contexte

ELDA a réalisé une annotation manuelle de plus de 12 000 énoncés puis envoyé à chacun des participants un échantillon de 3 000 énoncés tests. Les performances des différents systèmes ont été évaluées en fonction de la conformité de leurs sorties à l'annotation sémantique de référence. Selon une terminologie classique, les erreurs ont été classées en trois catégories : ***substitution/suppression/insertion (S/D/I)*** par rapport aux segments sémantiques de l'annotation. Afin d'éviter le report d'une insertion ou d'un délétion sur le reste des annotations de l'énoncé (ces types d'erreurs entraînent un décalage entre les annotations du système et celles de référence), un alignement des annotations par programmation dynamique a été réalisé.

L'exemple suivant est extrait du fichier d'évaluation de LOGUS ; REF correspond à l'annotation attendue. HYP est l'annotation proposée par LOGUS et obtenue par projection du graphe :



Il y a 3 segments considérés comme corrects. Le segment ?:OBJET:PAIEMENT-MONTANT est oublié (D : délétion) et le segment +:LIENREF-COREF:SINGULIER est remplacé (S : substitution) par +:OBJET:PAIEMENT-MONTANT-CHAMBRE. Ces erreurs sont essentiellement dues à des artifices utilisés pour respecter l'ordre des segments que la représentation produite par LOGUS ne permet pas a priori de respecter.

« alors l' hôtel avec piscine à combien est- il »

Scores: (#C #S #D #I) 3 1 1 0

REF: +:lienref-coref:singulier +:objetbd:hotel +:hotel-services:piscine ?:OBJET:PAIEMENT-MONTANT +:LIENREF-COREF:SINGULIER

HYP: +:lienref-coref:singulier +:objetbd:hotel +:hotel-services:piscine ********** +:OBJET:PAIEMENT-MONTANT-CHAMBRE

Eval: D S

Dans le deuxième exemple, le problème posé par les projections d'une forme de représentation sémantique en une autre forme apparaît encore plus clairement. Pour LOGUS, l'expression « *à quatre-vingt-dix euros maximum* » correspond à un unique segment sémantique strictement équivalent à « *maximum quatre-vingt-dix euros* ». L'ordre des mots ne peut être reconstitué à partir de la représentation sémantique rendue par LOGUS. Cette inversion « vaut » deux fautes : une de type D (délétion) et une autre de type S (substitution).

« alors je souhaite réserver deux chambres simples à quatre-vingt-dix euros maximum hein »

Scores: (#C #S #D #I) 5 0 1 1

REF: +:command-tache:reservation +:nombre-chambre-reservation:2 +:chambre-type:simple ********** +:paiement-montant-entier-reservation-chambre:90 +:paiement-monnaie:euro +:COMPARATIF-PAIEMENT:INFERIEUR

HYP: +:command-tache:reservation +:nombre-chambre-reservation:2 +:chambre-type:simple +:COMPARATIF-PAIEMENT:INFERIEUR +:paiement-montant-entier-reservation-chambre:90 +:paiement-monnaie:euro **********

Le troisième exemple fait apparaître encore plus clairement l'écart qui peut exister entre l'annotation sémantique de référence, proche de la forme linguistique, et les représentations sémantiques internes des systèmes :

« … vingt-quatre chambres simples et Paris dix-huitième »

Scores: (#C #S #D #I) 9 2 0 0

REF: +:nombre-chambre:24 +:chambre-type:simple +:CONNECTPROP:ADDITION +:localisation-ville-hotel:paris +:LOCALISATION-ARRONDISSEMENT-HOTEL:18

HYP: +:nombre-chambre:24 +:chambre-type:simple +:LOCALISATION-ARRONDISSEMENT-HOTEL:18 +:localisation-ville-hotel:paris

Outre le problème lié à l'ordre ville-arrondissement, apparaît ici le problème des connecteurs. Dans un système tel que Logus, les formes syntaxiques de la coordination servent à l'analyse de l'énoncé, mais elles peuvent ne pas apparaître dans la représentation

sémantique rendue. En revanche, on peut trouver dans la formule logique construite par le système des coordinations logiques qui ne sont pas présentes dans la forme linguistique initiale de l'énoncé.

4. DISCUSSION ET CONCLUSION

Les problèmes relatifs à la campagne d'évaluation proviennent soit du corpus, soit de la forme de l'annotation retenue.

4.1. Problèmes liés au corpus

Le corpus MEDIA peut paraître bien modeste aux yeux des spécialistes des corpus. Pour les développeurs de systèmes de DOHM il est néanmoins extrêmement intéressant car les corpus de dialogue de ce type en français sont très rares. Sous sa forme retranscrite, il présente néanmoins deux gros défauts.

- Le système de dialogue utilisé est très directif : le système interroge, l'utilisateur répond. Aussi, les énoncés sont-ils souvent très courts. Les énoncés longs sont presque toujours liés à un ou deux contenus sémantiques déterminés : description des conditions d'un séjour ou description d'une suite de séjours (cf. exemples donnés dans le 3.2.2.).
- Par ailleurs, la forme retranscrite qui a été utilisée pour l'évaluation ignore le problème majeur pour les systèmes de compréhension automatique que sont les erreurs du système de reconnaissance de la parole.

4.2. Problèmes liés à la représentation sémantique choisie

Le problème essentiel posé par la campagne d'évaluation hors-contexte de MEDIA provient du choix de la représentation sémantique. Le choix d'une représentation sémantique commune à laquelle tous les systèmes doivent se conformer est déjà un problème en soi car il est évident que les systèmes ne sont pas égaux devant la traduction à faire : plus grande est la distance entre leur propre représentation sémantique et celle qui est demandée, plus difficile est l'exercice. Pour LOGUS par exemple, il s'agissait d'un exercice très périlleux : les deux tiers des erreurs relevées proviennent de cette projection.

D'autre part, quoique classique, la représentation sémantique choisie est assez contestable. Il s'agit d'une représentation linéaire (plate) où la structure qui lie les différents segments n'apparaît pas directement : cette insuffisance est compensée par la complexification des attributs (par exemple *paiement-montant-entier-réservation-chambre*). La propagation des modes interrogatif et négatif est également assez difficile à maîtriser. Comme les exemples le montrent, l'alignement strict entre les segments et l'annotation pose un problème pour l'évaluation. En bref, le faible pouvoir expressif de cette représentation induit des difficultés de mise en œuvre. Pour l'application de réservation hôtelière, son expressivité semble à peu près suffisante. Il serait néanmoins intéressant de faire des « traductions inverses » pour savoir si l'annotation sémantique permet toujours de reconstituer le sens de l'énoncé initial.

4.3. Conclusion

Malgré ces réserves, l'expérience MEDIA demeure très intéressante. Outre l'intérêt du corpus en soi, se confronter au problème d'une annotation sémantique oblige à se poser de nombreuses questions. Les manuels rédigés pour l'annotation ont contraint à rédiger des

règles précises sans pour autant épuiser les sujets de débat : les très nombreuses discussions entre les participants montrent la complexité du problème de la représentation du sens et donc celui de la compréhension automatique. Ces confrontations et échanges d'idées sont, nous semble-t-il, indispensables pour faire progresser un domaine aussi complexe.

5. RÉFÉRENCES

Aït-Mokhtar S., Chanod J.-P., Roux C. 2002. 'Robustness beyond Shallowness : Incremental Deep Parsing' in *Natural Language Engineering*, volume 8, p. 121-144.

Allen J., Byron D. K., Dzikovska M., Galescu G., Stent A. 2001. 'Towards Conversational Humen-Computer Interaction' in *AI Magazine*, 22(4), p. 27-37.

Antoine J. Y., Bousquet C., Goulian J., Kurdi M. Z., Rosset S., Vigouroux N., Villaneau J. 2002. 'Predictive and Objective Evaluation of Speech Understanding : the challenge evaluation campaign of the I3 speech workgroup of the French CNRS' in *Proceedings of the LREC 2002*, Las Palmas, Spain.

Austin J-L. 1962. *How to do things with words*, Oxford.

Bonneau-Maynard H., Rosset S., Ayache C., Kuhn A., Mostefa D., MEDIA consortium. 2005. 'Semantic Annotation of the French Media dialog corpus' in *Proceedings of Eurospeech 2005*, Lisboa, Portugal.

Devillers L., Maynard H., Paroubek P. 2002. 'Méthodologies d'évaluation des systèmes de dialogue parlé : réflexions autour de la compréhension' in *Traitement automatique des langues (TAL)*, Volume 43, n°2, p. 155-184.

Glass J. 1999. 'Challenges for Spoken Dialogue Systems', in *Proceedings IEEE ASRU Workshop*, Keystone, Colorado, USA

Lamel L. , Bennacef S., Bonneau-Maynard H., Rosset S., Gauvain J-L. 1995. 'Recent developments in spoken language systems for information retrieval', in *Proceedings of the European Conference on Speech Technology (Eorospeech'95) Communication*, p. 207-210.

Minker W. 1998. 'Evaluation Methodologies for Interactive Speech Systems', in *Proceedings of the LREC'98*, Granada, Spain.

Sowa J. 2000. 'Knowledge Representation'. Brooks/Cole Thomson Learning, USA

Vanderveken D. 2001. 'Essays in Speech act Theory', in *Universal Grammar and Speech act Theory*. John Benjamin, Amsterdam Philadelphia, D. Vanderveken and Susumu Kubo Edition, p. 25-62.

Villaneau J., Ridoux O., Antoine J-Y. 2004. 'LOGUS : compréhension de l'oral spontané' in *Revue d'intelligence artificielle (RIA).* Volume 18. n°5-6. Hermes. Lavoisier.

6. REMERCIEMENTS

Merci à tous les membres du consortium MEDIA.

BASES DE DONNEES ET TRADUCTIONS : ASPECTS QUANTITATIFS ET QUALITATIFS

Yvon Keromnes
Université de Metz & Atilf-CNRS Nancy

1. INTRODUCTION

Nous nous proposons dans cet article d'aborder d'un point de vue méthodologique l'utilisation de bases de données en linguistique et en théorie de la traduction. En matière d'études de corpus, la tendance actuelle est à l'exploitation extensive de corpus étendus, ce qui correspond à une logique à la fois pratique et théorique. Pratique tout d'abord, puisque la tendance reflète l'évolution des logiciels et matériels informatiques permettant un traitement systématique et toujours plus rapide de bases de données toujours plus importantes. Théorique aussi, parce que cette évolution recouvre des enjeux épistémologiques liés à la taille et à la variété des corpus, ainsi qu'aux instruments qui leur sont appliqués : les enseignements que l'on pourra tirer d'un corpus sont souvent liés à une taille critique de celui-ci ; ainsi peut-on lire dans la préface du *Collins Cobuild Dictionary*, que son équipe de lexicographes, qui dans les années 80 s'enorgueillissait de son corpus de 20 millions de mots, avait dû déchanter en constatant l'impossibilité de proposer une description adéquate de la totalité de la langue anglaise à partir d'un tel corpus. Un corpus de 200 millions de mots permettrait d'y parvenir.

Dès lors, la taille semble être le paramètre déterminant pour la recevabilité des corpus. Dans le domaine français, *Frantext*, qui atteint également 200 millions de mots dans sa version intégrale, constitue la ressource privilégiée pour les textes littéraires. Pour la presse, la référence est le journal *Le Monde*. Or, l'utilisation de ces deux ressources n'est pas sans poser des problèmes, en premier lieu ceux de l'uniformité et de la représentativité : dans quelle mesure les textes les plus classiques reflètent-ils la langue française ou même simplement la littérature ? Et si la langue journalistique possède ses spécificités, *Le Monde* est-il prototypique à cet égard ? Enfin, que peuvent nous apprendre des statistiques sur ce type de corpus, outre la confirmation de leur homogénéité ?

Pour certains, la solution à ces problèmes sera dans la course en avant, l'application d'outils mathématiques toujours plus sophistiqués à des corpus toujours plus grands. Dans cette perspective, seul le web lui-même serait bientôt un corpus digne d'intérêt... à condition qu'il poursuive une croissance accélérée.

2. CORPUS ET TRADUCTION

Quelle est la pertinence de ces problèmes… et de leurs solutions pour des études de corpus ayant pour but de théoriser la traduction ? G. Bourquin (1984) a pu écrire à ce sujet :

> « Pour appréhender clairement les variables interlinguistiques pertinentes pour la traduction, une étude sur un corpus très étendu est nécessaire ».

Pourtant, sans vouloir rejeter la notion de validité liée à la taille du corpus, c'est une approche très différente que nous voulons proposer ici, avec une base de données élaborée

dans le cadre de notre thèse (Keromnes, 2000) consacrée au fonctionnement des formes verbales dans la narration en allemand, en anglais et en français, et qui a la particularité de permettre la comparaison entre un texte original allemand (*La Métamorphose*, de F. Kafka) et huit traductions publiées de cette nouvelle, quatre traductions en anglais et quatre en français, soit au total, si l'on associe l'original à l'ensemble paraphrastique, neuf versions d'un même texte. La conception de cette base de données de taille très modeste (environ 190 000 mots pour le corpus non balisé) découle des principes et objectifs exposés ci-dessous.

Le premier principe est notre approche paraphrastique de la traduction. L'opération de traduction articule quatre dimensions, celles de la langue source, du texte source, de la langue cible et du texte cible, ce dernier étant notablement instable[1] ; c'est en faisant reposer essentiellement la traduction sur la seule mise en rapport de deux systèmes linguistiques que Vinay & Darbelnet (1958 : 24) pouvaient écrire :

> « Il est permis de supposer que si nous connaissions mieux les méthodes qui gouvernent le passage d'une langue à l'autre, nous arriverions dans un nombre toujours plus grand de cas à des solutions uniques. »

Aujourd'hui, au contraire, nous dirons qu'à partir d'un texte original, la pluralité des traductions ne doit pas être considérée comme une exception, mais comme la règle. Et c'est cette pluralité qui constitue notre objet d'étude. Nous rejoignons sur ce point M. Ballard (1995 : 11) :

> « Une étude traductologique se doit, à nos yeux, de tenir compte des manières de traduire, qui sont diverses et parfois en désaccord ».

L'articulation des quatre dimensions de la traduction ne peut selon nous être véritablement appréhendée dans une problématique de la paraphrase (cf. C. Fuchs, 1975), non seulement dans une perspective interlinguistique, mais également intralinguistique ; le fait que l'on ne puisse aujourd'hui prétendre, de façon générale, à l'existence d'une solution unique en matière de traduction impose de prendre en compte les variantes éventuellement contradictoires, mais le texte source doit lui-même être mis en regard des différentes variantes qui peuvent lui être opposées[2]. Au centre des ensembles paraphrastiques intralinguistiques ou extralinguistiques se trouve le texte source, en regard duquel chaque paraphrase est produite, ce qui a pour effet de limiter fortement le nombre de paraphrases possibles[3]. D'où l'idée d'une base de données permettant de comparer systématiquement plusieurs traductions, en l'occurrence dans deux langues cibles, d'un même texte. Et notre conception de la paraphrase nous donne à penser que la comparaison des ensembles paraphrastiques anglais-français sera aussi utile que les comparaisons allemand-anglais et allemand-français, l'original allemand constituant dans le premier cas un *tertium comparationis* implicite. Autrement dit, dans cette base de données, le facteur d'extensivité, en tant que critère de validité des études de corpus, cède la place à un double facteur de complexité : au texte source correspondent quatre textes cibles dans deux langues cibles, eux-mêmes objets de comparaison.

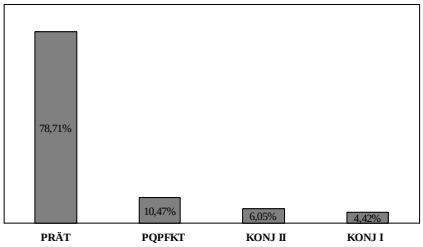
Notre objet d'étude étant les formes verbales dans la narration et la traduction, la base de données permettant cette étude doit répondre à deux objectifs antinomiques : permettre d'observer la variabilité des représentations sémantiques à partir d'un référent supposé stable et, d'autre part, conserver la structure de surface, dans laquelle plusieurs facteurs touchant à la textualité sont en jeu (ordre des termes, cohésion, cohérence…). D'où notre décision de retenir finalement le format du tableur Excel pour cette base de donnée. Texte

original et traductions y sont découpés en *équivalents propositionnels* (ci-après EP) selon une conception de la *proposition* exposée dans François & Keromnes (1995 : 45-46) et qui se situe à mi-chemin entre celle de Kintsch & Van Dijk (1975 : 99), et celle de la grammaire traditionnelle : nous proposons un découpage propositionnel reposant sur un élargissement de l'acception traditionnelle de « proposition » à tous les syntagmes (nominaux, adjectivaux, adverbiaux, participiaux) à fonction prédicative comportant un complément ou une détermination adverbiale et/ou étant introduits par une préposition non régie. [A person [knowing a lot]P] vs [A knowing person] : 'knowing a lot' étant reconnu comme syntagme dépendant, le premier exemple est compté comme EP alors que le second ne l'est pas. De la sorte, nous sommes à même d'observer la traduction de formes verbales finies par d'autres formes verbales finies que la traduction de formes verbales non finies, voire de formes non verbales, par des formes verbales finies, et inversement.

Enfin, d'un point de vue méthodologique, nous pensons qu'une base de données telle que nous l'envisageons doit répondre à un *critère de complétude* en présentant l'intégralité d'un texte et des traductions dont nous disposons. En effet, d'une part, pour le traducteur, le texte à traduire constitue en soi une unité et on ne peut prétendre véritablement cerner le fonctionnement des formes verbales entre un texte et ses traductions si l'on se limite à étudier des fragments : comme l'a montré Weinrich (1973), certaines formes verbales sont plus représentées à certaines étapes d'un récit, en français par exemple, des imparfaits seront plus nombreux en début ou en fin de récit, des passés simples dans le déroulement[4]. Par ailleurs, en présentant les traductions dans leur intégralité, nous nous interdisons de préjuger de la recevabilité des observations que nous pourrons faire sur elles : pour juger d'une tendance propre à un système linguistique particulier comparé à un autre, ou pour juger de la pertinence d'une traduction donnée d'une forme verbale, il faut disposer des différents choix effectivement faits en discours. Sans quoi la tentation serait trop grande de ne retenir que les tendances qui correspondent à ce que l'on veut démontrer.

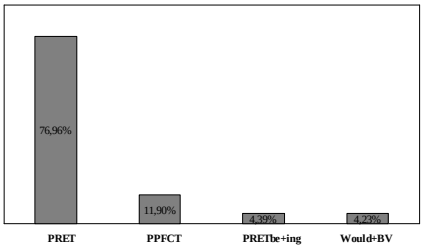
3. EXTRACTIONS ET DENOMBREMENTS

Comme pour les corpus de taille plus importante, l'exploitation de celui-ci revêt en premier lieu une dimension quantitative. Dans Excel, chaque texte, placé dans une colonne, conserve sa lisibilité, tandis que chaque ligne fait correspondre à un EP de l'original allemand quatre EP en français ou en anglais. Les EP sont pourvus de différentes étiquettes syntaxiques et sémantiques, dont la principale, l'indication de la marque de temps/mode/aspect du morphème verbal, mais aussi les numéros de paragraphes, phrases et propositions, le niveau de subordination, catégories et fonction syntaxiques des EP.... La base de données ainsi constituée intègre donc à la fois le corpus et les instruments de son exploitation, ce qui facilite la reproductibilité des observations : en mode filtre, chaque tête de colonne présente un menu déroulant affichant le contenu de la colonne ; par la sélection d'une ou plusieurs valeurs de ce menu, nous pouvons en extraire instantanément toutes les occurrences de ces valeurs dans la colonne. Ainsi, par la sélection successive des différents marqueurs de temps, mode ou aspect dans la colonne appropriée (TMA) correspondant au texte original, nous pouvons représenter la répartition des formes verbales dans ce texte.

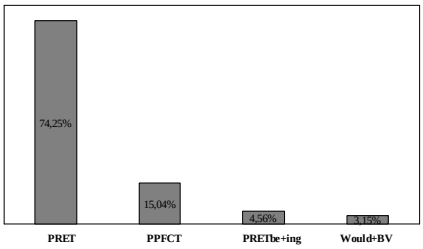


Graphique 1 : répartition des formes verbales finies principales[5] dans die Verwandlung (Präteritum, Plusquamperfekt, Konjunktiv II, Konjunktiv I)

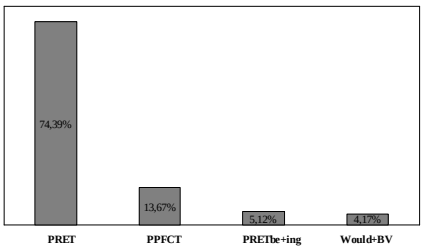
De même pour chacune des traductions (ici les traductions anglaises) :



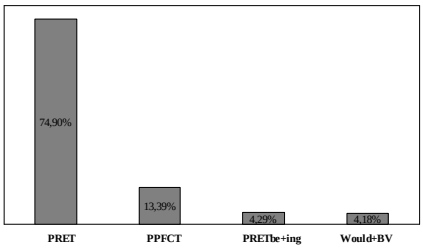
Graphique 2 : traduction de Muir & Muir



Graphique 3 : traduction de Underwood



Graphique 4 : traduction de Pasley



Graphique 5 : traduction de Corngold

Et alors que l'on peut s'attendre à ce que chaque traducteur tende à se distinguer de ses prédécesseurs, l'utilisation des morphèmes verbaux dans les traductions se montre relativement homogène. Si nous appliquons à présent le même filtre TMA dans les quatre traductions, nous obtenons la répartition des formes verbales faisant l'objet de décisions unanimes. Ce taux d'accord moyen nous donne une nouvelle indication sur l'emploi de ces formes verbales :

Formes verbales finies	Taux d'accord moyen
PRET	77,70 %
PPFCT	58,57 %
PRETbe+-ing	34,54 %
Would+BV	37,56 %

Tableau 1 : taux d'accord dans le choix des formes verbales en anglais

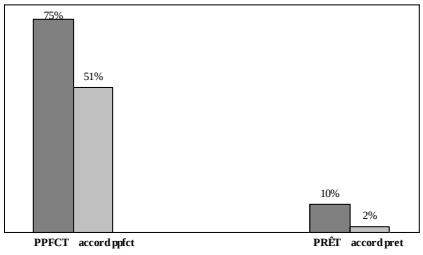
Nous voyons que si le choix d'un *prétérit* est le plus souvent unanime, celui d'un *pluperfect*, deuxième forme verbale la plus représentée, offre matière à une variation bien plus grande, et que les quatre traductions ne se retrouvent qu'à peu près une fois sur trois pour l'emploi des deux autres formes verbales. Ce taux d'accord est révélateur de l'existence de contraintes linguistiques plus ou moins fortes pesant sur le choix des traducteurs[6].

Si nous appliquons un filtre TMA dans l'original pour extraire, par exemple, les occurrences du *Plusquamperfekt*, deuxième forme verbale la plus représentée, puis un deuxième filtre aux différents marqueurs TMA d'une traduction, nous obtenons la répartition des traductions du *Plusquamperfekt* dans ce texte cible, aussi bien dans leurs aspects quantitatifs (nombre d'occurrrences de *pluperfect simple* ou *périphrastique*, de *prétérit* etc. pour une traduction en anglais) que qualitatifs (nombre de possibilités de traduire différemment un prétérit allemand) ; et là encore, nous pouvons comparer les traductions entre elles.

	Muir & Muir	Underwood	Pasley	Corngold
PPFCT	143	163	171	173
PRET	50	19	23	22
Would+BV	1	0	0	0
Would+Have+ -en	2	2	2	1
PPFCTbe+-ing	4	2	3	5
PRETbe+-ing	0	0	2	1

Tableau 2 : traductions du Plusquamperfekt (212 occurrences) en anglais

Nous n'indiquons ici que les traductions par des formes verbales finies (il manque donc les traductions par des formes verbales non finies, par des formes non verbales ou encore les non-traductions) mais nous voyons déjà d'une part la diversité qualitative, et d'autre part la concentration quantitative des traductions sur deux formes verbales. Là encore, si nous comparons les moyennes d'occurrence de ces deux formes verbales à leur taux moyen d'accord, nous obtenons une représentation différenciée de cet emploi :



Graphique 6 : traduction du Plusquamperfekt en anglais, emplois comparés des pluperfect et prétérit et taux d'accord moyens

Ce rapprochement de la répartition des traductions du *Plusquamperfekt* et des taux d'accord dans les traductions anglaises fait apparaître une forte contrainte pour les traductions par des *pluperfects* alors que l'emploi des *prétérits*, qui correspond tout de même à 10 % des traductions du *Plusquamperfekt*, fait partie des choix non contraints.

L'utilisation de filtres multiples permet enfin de comparer la variabilité des traductions selon des critères syntaxiques (nature des propositions, degré d'enchâssement,...) et textuels (début ou fin de paragraphe,...).

D'un point de vue linguistique, une étude quantitative de la répartition des formes verbales permet de mieux jauger des proportions de l'identique et du différent ; nous l'avons vu dans le cas de l'anglais, mais entre l'anglais et le français, on constatera également qu'un *prétérit simple* pourra correspondre non seulement à un *imparfait* ou à un *passé simple*, mais aussi à un *plus-que-parfait*, un *conditionnel*, un *subjonctif*, et même un *présent*, un *infinitif* ou un *participe présent* ; en revanche, dans 90 % des cas, en face du *prétérit simple*, ce sera bien un *imparfait* ou un *passé simple* que nous trouverons.

4. CONCLUSION

Nous avons voulu montrer, à l'encontre d'une tendance actuelle dans les études de corpus, l'intérêt linguistique et traductologique d'une base de données pour laquelle le critère de validité instrumentale ne réside pas dans la taille, mais dans la possibilité d'observer, sur des ensembles paraphrastiques et selon de multiples paramètres, la variabilité, tant qualitative que quantitative, des traductions, ainsi que la possibilité d'extraire à chaque fois les occurrences correspondant à des emplois linguistiquement contraints ou au contraire stylistiquement libres, deux possibilités associées à ce que nous avons appelé *critère de complétude*. Les limites de la démarche sont claires : il s'agit d'un texte littéraire particulier, dans lequel nos observations ne portent que sur le discours narratif, et les traductions observées sont uniquement orientées de l'allemand vers l'anglais et vers le français (c'est-à-dire vers une explicitation du marquage aspectuo-temporel). Pourtant, nous pensons que la description exhaustive des 2 380 équivalents propositionnels de l'original et de leurs traductions, la multiplication des perspectives que crée la mise en correspondance de 9 variantes d'un même texte sur trois langues compensent largement ces limites, pour faire de cette base un instrument tout à fait pertinent de validation... ou de falsification théorique.

5. REFERENCES

5.1. Corpus

5.1.1. TEXTE ORIGINAL :

Kafka F. 1983 (1ère publication : 1915). `'Die Verwandlung' in *Das Urteil*, Hamburg :Fischer.

5.1.2. TRADUCTIONS ANGLAISES :

Corngold S. 1996 (copyright traduction : 1972). *The metamorphosis*, New York, London : a Norton critical edition.

Muir W. & Muir E. 1995 (copyright traduction : 1949). *Kafka, the complete short stories*. London : Minerva.

Pasley M. 1992. *The Transformation and Other Stories*. London : Penguin.

Underwood J. A. 1991 (copyright traduction : 1981). *Franz Kafka Stories 1904-1924*. London : Cardinal.

5.1.3. TRADUCTIONS FRANÇAISES :

David C. 1990 (copyright traduction : 1989). *La métamorphose et autres récits*. Paris : Gallimard.

Lortholary B. 1988. *La métamorphose, Description d'un combat*. Paris : GF-Flammarion.

Vergne-Cain B. & Rudent G. 1988. *La métamorphose*. Paris : le Livre de Poche.

Vialatte A. 1955. *La métamorphose*. Paris : Gallimard.

5.2. Linguistique et traductologie

Ballard M. 1987. *La traduction : de l'anglais au français*. Paris : Nathan.

Ballard M. (ed). 1995. *Relations discursives et traduction*. Lille : Presses universitaires de Lille.

Bourquin G. 1984. *Semantic Relations, Modality, Time-Tense. A resolution study on French*, Document C.E.E., Eurotra Report ETL-F-1 (4), Luxembourg, 1984, 220 p.

Demanuelli C. & Demanuelli J. 1991. *Lire et traduire : anglais-français*. Paris : Masson.

François J. & Keromnes Y. 1995. 'De la fidélité en traduction littéraire : esquisse d'une méthode d'évaluation comparative', in *Lectures, Hommage à Geneviève Hily-Mane*, Presses Universitaires de Reims, p. 41-68.

Fuchs C. 1994. *Paraphrase et énonciation*, Paris : Ophrys.

Keromnes Y. 2000. *Formes verbales, narration et traduction : étude d'une nouvelle de F. Kafka et de quatre traductions de cette nouvelle en anglais et en français*, Thèse de doctorat, Université Nancy 2.

Keromnes Y. 2005. 'Equivalences paraphrastiques : une question de perspective', in *Recherches en Linguistique et Psychologie cognitive* 21, Presses Universitaires de Reims. p. 241-261.

Kintsch W. & van Dijk T. 1975. 'Comment on se rappelle et on résume des histoires', in *Langages* 40, Paris : Larousse. p. 98-116.

Mel'cuk I. 1993. *Cours de morphologie générale*, vol. 1, Presses de l'Université de Montréal, CNRS Editions.

Weinrich H. 1973. *Le Temps* (trad. française M. Lacoste), Paris : le Seuil.

6. NOTES

1. Cf. C. Demanuelli & J. Demanuelli (1991 : 7) ; également M. Ballard (1987 : 45), qui explique par l'instabilité du texte à produire la variabilité de l'unité de traduction.

2. Pour plus de détails sur cette approche paraphrastique de la traduction, cf. Keromnes (2000, 2005).

3. Nous sommes donc ici à l'opposé de l'approche des paraphrases en termes de synonymie relative de I. Mel'cuk (1993 : 44), qui dans son modèle de génération paraphrastique, parvient en théorie à un ensemble de 50 millions d'éléments à partir d'une simple phrase extraite d'une dépêche d'agence de presse.

4. L'importance de cette prise en compte de la dimension textuelle ne se limite d'ailleurs pas à l'étude des formes verbales. Une étude du connecteur *and* que nous avons mené sur ce même corpus fait apparaître des emplois tout à fait particuliers en fin de paragraphe, de section et de la nouvelle elle-même.

5. Les formes verbales représentant moins de 1 % de l'ensemble des formes verbales finies (par ex. le présent, 6 occurrences, 0,3 % de l'ensemble) ne sont pas représentées sur ce graphique.
6. On pourrait penser que le taux d'accord est d'autant plus fort que la forme verbale est fortement représentée, or il n'en est rien. Ainsi, dans les traductions françaises, alors que le *passé simple* est sensiblement moins employé que *l'imparfait*, le taux d'accord moyen est de 72,50 % pour le premier et de 62,85 % pour le second : les contraintes aspectuelles liées à la progression narrative sont donc bien plus fortes pour le *passé simple* que pour l'*imparfait*.

APPRÉCIATIONS SUR LA FIABILITÉ D'UN ÉTIQUETAGE MORPHOSYNTAXIQUE GÉNÉRÉ AUTOMATIQUEMENT ET PRÉSENTATION DU TRAVAIL LIÉ A SON ENRICHISSEMENT

Vassil Mostrov
UMR Savoirs, Textes, Langage (UMR 8163 du CNRS) - Université Lille 3

1. INTRODUCTION

Ma contribution s'articulera autour de deux objets principaux : (i) à partir de l'utilisation de la fonction « étiqueter un texte » de Cordial Analyseur 7.0, je pointerai différents types d'erreurs analysées en contexte, afin de pouvoir formuler une appréciation sur la fiabilité de l'étiquetage automatique et (ii) à partir d'imprécisions observées dans l'étiquetage et des ambiguïtés qu'elles génèrent pour les traitements à venir, je montrerai la pertinence de mon travail d'enrichissement des informations morphosyntaxiques données dans les étiquettes dites de « type grammatical » dans la terminologie de Cordial Analyseur.

Avant de développer ces deux points, je me propose de présenter brièvement le corpus de travail sur l'étude duquel sera basée l'analyse des problèmes discutés dans le présent article. Ce corpus a été élaboré à L'UMR Silex/STL1, Université de Lille 3, dans le cadre du projet OuRAL2. C'est un corpus d'oral transcrit ayant trait à la gestion locative de biens immobiliers. Il a été constitué selon les étapes suivantes : a) réalisation d'entretiens auxquels ont participé des propriétaires bailleurs et des locataires ; b) numérisation des enregistrements de ces entretiens ; c) transcription avec le logiciel Transcriber 1.4. en alignant chaque tour de parole transcrit à sa source sonore et en notifiant quelques événements remarquables3 ; d) mise en conformité des transcriptions avec une DTD simplifiée adaptée aux exigences des posttraitements à suivre ; e) traitement des fichiers XML ainsi obtenus avec le logiciel Cordial Analyseur 7.0 (désormais abrégé CA7.0) afin d'obtenir l'étiquetage morphosyntaxique des propos transcrits ; f) conversion des fichiers tabulés produits par CA7.0 en fichiers XML où chaque mot est balisé comme item, avec en attributs les informations liées à l'étiquetage morphosyntaxique.

Les deux dernières étapes méritent d'être mieux décrites. Nous avons demandé à CA7.0 de nous fournir deux versions d'étiquetage morphosyntaxique : avec et sans reconnaissance des expressions. Dans la première version, les expressions sont reconnues sans être analysées. Nous avons alors utilisé la deuxième version pour enrichir la première, afin d'en créer une troisième où les expressions sont reconnues globalement et les éléments qui les composent étiquetés :

```
<EXPRESSION lemme="tacite reconduction" typegram="NCFS">
<ITEM lemme="tacite" typegram="ADJIGS">tacite</ITEM>
<ITEM lemme="reconduction" typegram="NCFS">reconduction</ITEM>
</EXPRESSION>
```

2. PRÉSENTATION TYPOLOGIQUE D'UNE SÉLECTION D'ERREURS D'ÉTIQUETAGE

Avant de présenter les types d'erreurs générées par CA7.0, je me propose d'évoquer deux facteurs principaux qui contribuent à expliquer l'existence de ces erreurs. Le premier concerne le mode de fonctionnement de l'étiqueteur en question : CA7.0 construit ses analyses phrase par phrase5. Pour lui, une fin de phrase est mécaniquement identifiée par la présence d'une ponctuation forte. Si, dans le texte fourni en entrée, il n'y a pas de ponctuation, comme cela est le cas dans le corpus d'oral transcrit6 développé dans le cadre du projet OuRAL, CA7.0 s'arrête à chaque fin d'énoncé et, pour les plus longs, considère qu'une nouvelle phrase débute tous les 127 mots. Le deuxième facteur concerne la nature de notre corpus. Du fait qu'il s'agit d'un corpus d'oral transcrit, les « faits de parole » sont abondants : répétitions, reformulations, inachèvements, amorces, ruptures syntaxiques, marques d'hésitation, etc., ce qui complique le travail de l'étiqueteur.

La plupart des erreurs qui m'occupent dans le cadre de ce travail sont liées à l'identification de la catégorie morphosyntaxique de mots homographes polycatégoriels. Il me semble utile d'introduire ici la terminologie qu'emploie Catherine Fuchs (1996 : 35) afin de distinguer deux types d'ambiguïtés : les « virtuelles » et les « effectives » : « Toute forme à laquelle peuvent être associées plusieurs significations en langue est virtuellement ambiguë si on la considère isolément, hors tout contexte d'emploi. Lorsque cette forme se trouve plongée dans un contexte linguistique, elle peut devenir univoque – on dit que l'ambiguïté virtuelle est levée par le contexte – ou être effectivement ambiguë – on dit alors qu'elle conserve son ambiguïté ».

Nous allons voir qu'une partie des ambiguïtés « virtuelles » pour l'humain sont « effectives » pour la machine. Ceci est vraisemblablement dû, au moins en partie, au fait que la machine n'est pas dotée de mécanismes appropriés pour l'interprétation du sens et qu'elle ne détient pas les connaissances sur le monde nécessaires. Mais le contexte syntaxique, analysé au moyen de règles distributionnelles adaptées, permet le plus souvent de lever ce type d'ambiguïtés : je le montrerai après avoir présenté et illustré différentes erreurs que j'ai relevées lors de la relecture de notre corpus morphosyntaxiquement annoté.

2.1. Erreurs dues à des « faits de parole »

Les « faits de parole » créent pour CA7.0 des contextes que j'appelle « virtuellement ambigus ». Ce sont des contextes où la distribution syntaxique est respectée entre des unités dont l'analyse morphosyntaxique est fautive. Autrement dit, il s'agit d'un choix erroné de catégorie pour deux ou plusieurs mots, la séquence d'étiquettes créant une structure « syntaxiquement » correcte (prep detdms ncms, pour le premier exemple ci-dessous).

2.1.1 RÉPÉTITION DE MOTS GRAMMATICAUX

(1) un avantage pourPREP leDETDMS pourNCMS l'DETDMS étudiantNCMS

(2) un salon et uneDETIFS unePIFS chambreNCFS

Dans (1), la deuxième unité pour est analysée comme ayant la catégorie nom. Ceci est dû d'abord au fait qu'à la forme le qui la précède, CA7.0 a attribué la catégorie déterminant – seule catégorie possible dans cet environnement (l'interprétation du le en tant que pronom personnel ne serait possible que s'il était suivi d'un verbe à l'infinitif ou d'un pronom préverbal mais pour ne peut pas être analysé ainsi). Dans la mesure où le est désambiguïsé de cette façon, et où il y a en français une règle syntaxique qui dit qu'un

déterminant introduit obligatoirement un constituant nominal, CA7.0 a choisi d'attribuer la catégorie nom à l'unité pour – cette dernière unité ayant potentiellement les catégories nom ou préposition dans le dictionnaire interne du logiciel. Quant à (2), La première occurrence de une a reçu la valeur de déterminant puisque c'est le premier candidat déterminant dans le syntagme fautivement repéré. La deuxième occurrence de une est ensuite analysée comme pronom indéfini féminin singulier puisqu'il est précédé d'un déterminant. Ceci nous laisse penser que CA7.0 fait ses analyses syntagme par syntagme sans revenir sur ses analyses antérieures : puisque l'étiqueteur a analysé une fois la succession « une une » comme étant un syntagme nominal du type déterminant-déterminé, il ne revient pas en arrière pour réattribuer une valeur à la deuxième occurrence de une au moment où il voit que celle-ci précède un nom (ici chambre) qui doit normalement avoir sa détermination.

2.1.2 RUPTURES SYNTAXIQUES OU REFORMULATIONS

(3) je loue euh àPREP uneDETIFS seuleNCFS enfinADV uneDETIFS seuleADJFS personneNCFS

(4) lesDETDIGP deuxADJNUM moyensADJMP lesDETDIGP deuxADJNUM sourcesNCFP que j'utilise

En (3), le choix de la catégorie nom pour le premier seule est le résultat de la prise en considération du seul environnement contextuel immédiat de cette unité : elle est précédée d'un déterminant et non suivie d'un nom. L'analyse de moyens en (4) reste par contre obscure : il semble que CA7.0 ait eu du mal à déterminer la valeur grammaticale de ce mot même dans des contextes univoques et syntaxiquement réguliers, ce qui va être confirmé par d'autres exemples (cf. 9 et 9' en 2.2).

2.2. Erreurs générées dans un contexte univoque

Il y a plusieurs cas où CA7.0 a généré des erreurs de catégorisation morphosyntaxique alors que le contexte environnant les unités fautivement analysées est univoque et que les structures syntaxiques où elles entrent sont régulières. La raison de l'existence de ce type d'erreurs reste pour moi obscure – du moins en essayant de les expliquer en me basant sur l'application de règles de distribution. Néanmoins, en regardant de près les exemples qui suivront, nous pourrons les regrouper en sous-ensembles cohérents :

(5) y'a un certain nombre d'équipements qui sont (quand même)SUB disponiblesADJIGP dans le logement

(6) votre remboursementNCMS mensuelNCMS aVINDP3S euVPARPMS une incidence

(7) (parce que)SUB bonADJMS je suis pas trèsADV bonINT dansPREP toutADJIND cePDS qui est

(8) je ne sais pas exactementADV siADV yPPER3S aVINDP3S quelque chose

(9) j'avais essayé de reprendre des études mais (faute de)PREP moyensADJMP j'ai pas pu

(9') on s'est concentrés sur des logements étudiants (parce que)SUB (faute de)PREP moyensNCMP euhINT nous étions à ce moment-là tous les deux étudiants

(10) c'est pas (quelque chose)PIMS queSUB vous possédiez déjà

(11) est-ce que ça vous aVINDP3S donnéVPARPMS droitADV àPREP des subventions

L'exemple (5) illustre un premier sous-type d'erreur. Il semble qu'ici le problème soit dû au codage de la catégorie morphosyntaxique dans le dictionnaire interne du logiciel : quand même est toujours étiqueté comme conjonction de subordination. Il est indéniable que cette unité peut très bien avoir cette valeur mais seulement à condition d'être suivie d'une proposition au conditionnel (= même si), ce qui n'est pas le cas du contexte en (5) qui impose à quand même une interprétation d'adverbe.

Les exemples de (6) à (9) correspondent à un deuxième sous-type d'erreurs. Il s'agit d'une analyse erronée de certaines unités malgré l'univocité du contexte et malgré le fait que la syntaxe de ces phrases ne présente aucune irrégularité : • en (6), mensuel dans cette position ne peut qu'être adjectif ; • en (7), la deuxième occurrence de bon reçoit également une valeur d'adjectif puisqu'il est (i) attribut du sujet et (ii) déterminé par l'adverbe très qui peut s'appliquer à des adjectifs, noms, participes passés et adverbes mais pas à des interjections ; • en (8), si introduit une complétive, à moins qu'on ne considère qu'on soit en présence d'une rupture syntaxique due aux « faits de parole », mais on a constaté dans le point précédent que CA7.0 est incapable de prendre en considération ce genre de phénomène ; • en (9), la locution prépositive faute de ne peut introduire qu'un nom (ou un SN) ou un verbe à l'infinitif ou encore un adverbe employé nominalement (ex : faute de mieux) mais pas un adjectif isolé (il est curieux de constater qu'en (9'), où le contexte syntaxique est le même à ceci près que moyens est suivi d'une marque d'hésitation et d'une parenthétique, CA7.0 a désambiguïsé correctement cette unité).

Les exemples (10) et (11) sont plus délicats à analyser puisque, à première vue, le contexte syntaxique permet immédiatement de lever l'ambiguïté mais qu'en réalité, c'est une analyse syntaxique passant par l'interprétation du sens qui est nécessaire pour résoudre de tels cas. C'est pour cette raison que je les considère comme faisant partie d'un troisième sous-type d'erreurs. En (10), que doit être interprété comme un pronom relatif ayant comme antécédent le pronom qui le précède immédiatement. Mais on sait par ailleurs qu'il existe des cas où que conjonction de subordination peut apparaître dans le même contexte syntaxique immédiat – par exemple après des structures impersonnelles du genre il est dommageNCMS que. Ces données créent un contexte immédiat ambigu pour la machine, qui ne saura pas décider s'il s'agit d'un pronom relatif ou d'une conjonction de subordination. En (11), CA7.0 n'a pas reconnu l'existence de l'expression verbale donner droit, où le nom est incorporé au verbe. Cette incorporation se manifeste syntaxiquement par l'absence de tout déterminant devant le nom. Or, généralement, l'unité droit dans un tel environnement syntaxique a la valeur d'adverbe – c'est ce qui se passe dans : viser droit, aller droit au but, marcher droit – et c'est certainement à partir de cette règle distributionnelle que CA7.0 a construit l'analyse de la collocation donner droit ADV. Des exemples comme (10) et (11) montrent qu'il y a des cas où le choix erroné d'une catégorie grammaticale est dû, entre autres choses, à l'absence d'une analyse sémantique. C'est précisément sur l'illustration de cette absence que va porter le point suivant.

2.3. Erreurs liées à la non-reconnaissance du sens

L'absence d'une interprétation sémantique semblable à celle que l'humain fait lors du décodage des énoncés linguistiques est, à mon sens, la cause principale de la plupart des erreurs présentées jusqu'ici. Nous avons constaté que l'application mécanique des règles de la distribution syntaxique n'est pas suffisante pour analyser correctement dans tous les cas la catégorie morphosyntaxique des unités analysées. Je me propose maintenant de présenter quelques types d'erreurs que je considère comme étant purement sémantiques, ce

qui ne veut pas dire que les erreurs de 2.1 et de 2.2 soient liées à des phénomènes exclusivement autres que d'ordre sémantique, dans un sens plus large du terme (peut-être que l'absence d'une analyse sémantique pertinente contribue d'une façon moins directe à ce qu'elles soient générées). Je considère que la seule non-reconnaissance du sens peut entraîner :

- L'analyse erronée d'une catégorie morphosyntaxique

(12) je suis propriétaire d'unDETIMS seulADJMS appartementNCMS

(13) unDETIMS moisNCMIN alors que c'est plutôt deuxADJNUM moisNCMIN qui sont pratiqués

Les exemples (12) et (13) montrent que CA7.0 n'est pas doté d'un mécanisme lui permettant de distinguer le un déterminant indéfini du un numéral. Si en (12) la présence du prédicat seul suffit à l'humain pour lever l'ambiguïté, en (13) cette ambiguïté est levée grâce à l'opposition un vs deux mois qui est perceptible en tenant compte de l'interprétation du sens de l'ensemble de l'énoncé.

- L'identification erronée de la valeur d'une désinence verbale

(14) j'aurais appelé les propriétaires (pour qu')SUB ils s'enPPER3S occupentVINDP3P

CA7.0 reconnaît les formes verbales comme étant au subjonctif seulement dans les cas où ces formes ne coïncident pas avec celles du présent de l'indicatif 7. S'il y a une telle coïncidence alors même que le contexte est explicite quant au choix subjonctif / indicatif présent, CA7.0 donne automatiquement l'étiquette « indicatif » à la forme verbale en question.

- L'identification erronée d'un lemme :

(15) on fait un inventaire avec la femmelemme="mari" de ménage de l'étage

(16) ma petite sœur lemme="frère" est venue à Lille

(17) je crois que je saurais lemme="saurer" me renseigner

(18) est-ce que t'lemme="te" avais fait d'autres démarches

Le lemme est la « forme canonique (…) d'un mot variable » (Petit Robert s.v. lemme 3) – c'est-à-dire la forme du mot tel qu'il se présente dans la nomenclature d'un dictionnaire (= entrée). Le lemme peut être considéré comme la forme du mot qui se rapproche le plus du concept puisqu'elle est hors emploi, et qu'elle couvre potentiellement tous les sens possibles du mot (si le mot est polysémique). Dans le discours au contraire, le mot s'ajuste à des références concrètes, ce qui se matérialise par les marques d'accord, de conjugaison, etc.

Nous avons demandé à CA7.0 de lemmatiser les mots susceptibles d'avoir un genre en nous donnant, si elle existe, leur forme du masculin. Pour les substantifs, cette possibilité existe quand la distinction masculin/féminin est basée sur la prise en considération du genre naturel des référents. Ainsi par exemple, la « variante féminine » de directeur est directrice. La formation du féminin dans ce cas est assurée par un procédé morphologique – la suffixation. En (15) et (16) en revanche, nous sommes en présence de l'application d'un procédé lexical, qui se traduit par l'existence de mots-couples : chaque référent, en fonction du genre, a son propre mot. Je considère que la lemmatisation des substantifs femme et sœur respectivement en mari et frère pose deux problèmes majeurs. Le premier est d'ordre sémantique : si sœur en (16) est indubitablement le correspondant féminin de frère, il n'en est rien de femme et de mari en (15), cette dernière unité étant la forme

masculine d'un des sous-sens de femme, à savoir celui d'épouse. Le deuxième problème est d'ordre lexical : puisqu'il s'agit de deux mots ayant des racines différentes pour chaque genre, ils ont une existence autonome, ce qui pourrait être corroboré par le fait qu'ils ont donné lieu à une polysémie qui leur est propre : ainsi sœur n'est pas uniquement le correspondant féminin de frère mais peut recevoir un sens non relationnel comme dans La sœur Thérèse. Une preuve de l'existence autonome des mots hétéronymiques (ou formes supplétives) pourrait être puisée dans le fait qu'à chaque mot du genre correspond un article dans les dictionnaires. Si l'on considère que le lemme d'un mot correspond à son entrée dans le dictionnaire, on comprend pourquoi le lemme de sœur est sœur, à la différence de celui de directrice par exemple, lequel mot apparaît sous l'adresse directeur puisqu'il est formé par suffixation et n'indique rien d'autre que le passage au genre naturel féminin.

Les exemples de (17) et de (18) présentent des erreurs de lemme plus mécaniques : en (17), c'est la non-reconnaissance du verbe savoir qui a provoqué la confusion entre la forme de la 1ère personne du conditionnel présent de ce verbe et la forme de l'imparfait de l'indicatif de la même personne du verbe saurer8. Notons que l'identification du verbe en (17) en tant que saurer est d'autant plus invraisemblable qu'il est suivi d'un pronom et d'un infinitif – environnement syntaxique impossible pour ce verbe. Enfin, l'exemple (18) montre que CA7.0 ne reconnaît pas le pronom personnel sujet de la 2e personne du singulier quand il est élidé devant une voyelle9 (ou un h muet). Ce phénomène, fréquent à l'oral et de l'ordre du familier, se rencontre à l'écrit dans le discours direct et n'est pas marginal. Des exemples de cette élision peuvent être trouvés déjà dans des textes du 17e siècle comme le prouve cette phrase de Molière, tirée du Petit Robert s.v. tu : « t'as la vue trouble (…) t'as la barlue [berlue] ».

2.4. Conclusion

Une partie des erreurs présentées et commentées en 2 relèvent clairement de facteurs liés aux spécificités de la production orale spontanée : l'absence de ponctuation dans les transcriptions ainsi que les abondants « faits de parole » ont perturbé l'analyse morphosyntaxique de CA7.0, logiciel mieux adapté au traitement de productions langagières écrites. En revanche, une partie des erreurs observées dans notre corpus (statistiquement le plus grand nombre) ne sont pas liées aux facteurs évoqués ci-dessus puisque les contextes syntaxiques où elles sont produites et les relations sémantiques au sein de ces contextes ne présentent aucune irrégularité, ce qui est montré par les exemples de 2.2 et de 2.3. La génération d'au moins une partie de ces dernières erreurs doit être probablement cherchée, comme le suggère Fuchs (1996), dans l'extrême rigidité et dans la surpuissance des systèmes d'analyse morphosyntaxique. Le système essaie d'appliquer une multitude de règles syntaxiques à chaque constituant afin d'en fournir une analyse correcte – plus on affine les règles, plus on a le risque d'une prolifération d'ambiguïtés artefactuelles. C'est ce qui semble s'être passé en (10) de 2.2 : la règle selon laquelle on peut avoir un nom devant la conjonction que dans certains cas semble l'avoir emporté sur la règle qui dit que que après un nom doit être interprété comme étant un relatif. Il est clair que pour faire une analyse morphosyntaxique correcte, l'analyseur automatique devrait intégrer des règles sémantiques complexes qui complèteraient les règles syntaxiques de distribution10.

3. ENRICHISSEMENTS APPORTES AUX ETIQUETAGES DE CORDIAL ANALYSEUR 7.0

Les étiquettes appelées « type grammatical » dans la terminologie de CA7.0 permettent de coder à la fois (i) l'appartenance des unités analysées à une catégorie morphosyntaxique donnée (nom, adjectif…) et (ii) les informations de catégorie grammaticale (souvent exprimées par des flexions) qui leur sont rattachées : genre, nombre, mode, temps et personne. Une étiquette morphosyntaxique de CA7.0 se présente de la façon suivante : tout d'abord, nous avons la composante « catégorie morphosyntaxique » (par exemple PPER pour les pronoms personnels) et ensuite, selon le cas, les spécifications relatives (i) à la personne (1, 2 ou 3) ; (ii) au genre (M, F ou IG pour les invariables en genre) ; (iii) au nombre (S, P ou IN pour les invariants en nombre) ; (iv) au mode (INFinitif, INDicatif, SUBjonctif, CONditionnel, IMPératif et PARticipe) ; et (v) au temps (Présent, Imparfait, PasséSimple, Futur). Toutes ces informations se trouvent concaténées sans espace les séparant.

À partir de la liste des étiquettes intégrées dans CA7.0, j'ai observé que des informations manquaient de façon irrégulière, notamment en matière de catégorie grammaticale, mais également concernant certains types de catégories morphosyntaxiques. Je propose maintenant de présenter sous forme de liste les principaux ajouts (correspondant aux codages donnés en gras dans la composition des étiquettes) que j'ai faits dans le cadre de l'annotation morphosyntaxique de notre corpus11 :

- **Pronoms personnels** : PPER3**IG**S (*lui* datif, *on*) ; PPER3**M**S (*il, lui, le, l'*) ; PPER3**F**S (*elle, la, l'*) ; PPER3**IG**P (*les, leur*) ; PPER3**M**P (*ils, eux*) ; PPER3**F**P (*elles*) ; PPER3**I** (*y, en, se*) ; PPER3**N** (*il* impersonnel).
- **Pronoms démonstratifs** : PD**M**S (*celui*[*-ci/-là*]) ; PD**F**S (*celle*[*-ci/-là*]) ; PD**N** (*ce, ça, ceci, cela, c'*) ; PD**M**P (*ceux*[*-ci/-là*]) ; PD**F**P (*celles*[*-ci/-là*]).
- **Pronoms relatifs** : PR**MP** ([*les-/des-/aux-*]*quels*) ; PR**FP** ([*les-/des-/aux-*]*quelles*).
- **Pronoms possessifs** : PP**MS1S** (*le mien*) ; PP**MS2S** (*le tien*) ; PP**MS3S** (*le sien)* ; PP**MS1P** (*le nôtre*) ; PP**MS2P** (*le vôtre*) ; PP**MS3P** (*le leur*) ; PP**FS1S** (*la mienne*) ; PP**FS2S** (*la tienne*) ; PP**FS3S** (*la sienne*) ; PP**FS1P** (*la nôtre*) ; PP**FS2P** (*la vôtre*) ; PP**FS3P** (*la leur*) ; PP**MP1S** (*les miens*) ; PP**MP2S** (*les tiens*) ; PP**MP3S** (*les siens*) ; PP**FP1S** (*les miennes*) PP**FP2S** (*les tiennes*) ; PP**FP3S** (*les siennes*) ; PP**IGP1P** (*les nôtres*) ; PP**IGP2P** (*les vôtres*) ; PP**IGP3P** (*les leurs*).
- **Adjectifs démonstratifs** : DETDEM**MS** (*ce, cet*) ; DETDEM**FS** (*cette*) ; DETDEM**IGP** (*ces*).
- **Adjectifs possessifs** : DTPS**IGS1S**[12] (*mon*) ; DTPS**IGS2S** (*ton*) ; DTPS**IGS3S** (*son*) ; DTPS**FS1S** (*ma*) ; DTPS**FS2S** (*ta*) ; DTPS**FS3S** (*sa*) ; DTPS**IGS1P** (*notre*) ; DTPS**IGS2P** (*votre*) ; DTPS**IGS3P** (*leur*) ; DTPS**IGP1S** (*mes*) ; DTPS**IGP2S** (*tes*) ; DTPS**IGP3S** (*ses*) ; DTPS**IGP1P** (*nos*) ; DTPS**IGP2P** (*vos*) ; DTPS**IGP3P** (*leurs*).
- **Adjectifs numériques** : ADJNUM**M** (*un*) ; ADJNUM**F** (*une*) ; ADJNUM**IG** (*deux…*) ; ADJORD**MS** (*premier…*) ; ADJORD**FS** (*première…*) ; ADJORD**MP** (*premiers…*) ; ADJORD**FP** (*premières…*) ; ADJORD**IGS** (*deuxième…*) ; ADJORD**IGP** (*deuxièmes…*).
- **Adjectifs interrogatifs** : ADJINT**MS** (*quel*) ; ADJINT**MP** (*quels*) ; ADJINT**FS** (*quelle*) ; ADJINT**FP** (*quelles*).

• **Adjectifs indéfinis** : ADJIND**MS** (*certain…*) ; ADJIND**MP** (*certains…*) ; ADJIND**FS** (*certaine…*) ; ADJIND**FP** (*certaines…*) ; ADJIND**IGS** (*autre…*) ; ADJIND**IGP** (*autres…*) ; ADJIND**I** (*chaque, plusieurs, beaucoup de…*) ; ADJIND**MIN** (*divers*) ; ADJIND**FIN** (*diverses*).

• **Articles** : DETDMS**A** (*du* [*de+le*], *au*) ; DETDIGP**A** (*des* [*de+les*], *aux*) ; DETIMS**PAR** (*du*) ; DETI**NEG** (*de* [*je ne mange pas de tomates*]) ; DETI**IGP** (*des, de* [*de belles roses*]).

• **Pronoms interrogatifs : PINTI** (*qui, que, quoi, qui est-ce qui, qu'est-ce qui, qui est-ce que, qu'est-ce que*) ; **PINTMS** (*lequel*) ; **PINTFS** (*laquelle*) ; **PINTFP** (*lesquelles*) ; **PINTMP** (*lesquels*).

En observant la liste ci-dessus, nous constatons d'abord que dans les étiquettes de CA7.0 relatives aux pronoms et aux adjectifs, il manque d'une façon irrégulière des informations liées aux catégories du genre, du nombre et de la personne. Par ailleurs, d'autres informations se sont révélées manquantes : (i) pour les pronoms personnels, une étiquette spécifique au il impersonnel ; (ii) pour les pronoms démonstratifs, une étiquette relative aux formes neutres ; (iii) pour les articles, des étiquettes permettant de coder les formes amalgamées, le partitif et le déterminant indéfini pluriel sous ses formes canonique et faible et (iv) des étiquettes relatives aux pronoms interrogatifs, catégorie non prise en compte dans le jeu de CA7.0.

3.1. Motivations des enrichissements effectués

Je me concentrerai sur quelques exemples pour montrer la pertinence d'une partie des ajouts que j'ai faits. La grande précision des informations codées dans les étiquettes de « type grammatical » me semble cruciale pour des requêtes ultérieures, basées sur la reconnaissance des unités d'un point de vue grammatical. Dans quelques cas, la précision des étiquettes actuelles est insatisfaisante :

- Nous avons besoin de coder le genre des pronoms personnels si, par exemple, nous voulons ne trouver dans le corpus, que les occurrences au masculin du pronom COD le, et ainsi filtrer les contextes où cette unité a comme antécédent un nom masculin (Je le connais [le voisin]) et ceux où elle renvoie à une proposition dénotant un fait (Je le sais [que la propriétaire viendra]). Puisque les fonctions syntaxiques des unités de notre corpus ne sont pas explicitées, il faudra intégrer dans la requête les contextes syntaxiques possibles de l'unité cherchée (dans ce cas, l'item immédiatement à gauche pourra être une préposition, un autre pronom personnel, un verbe à l'impératif ou une particule de négation).

- Le codage spécifique du pronom neutre il s'avère utile pour la recherche d'expressions impersonnelles (il est probable que…) ou de constructions dans lesquelles entrent des verbes impersonnels (il neige) afin de différencier ces cas de ceux où le sujet s'associe clairement à un référent masculin de la troisième personne du singulier (il pense que…).

- Il n'est pas sans intérêt de coder séparément les formes dites neutres des pronoms démonstratifs puisque certaines d'entre elles (notamment ceci, cela et ça) peuvent anaphoriser des antécédents dépourvus de genre et de nombre que sont les propositions (On fait l'état des lieux. Après ça on descend au secrétariat).

- Concernant les pronoms relatifs, dans le jeu d'étiquettes de CA7.0 sont codées uniquement les formes masculines et féminines du singulier, et les formes du pluriel sont étiquetées de la même façon que les formes invariantes en genre et en nombre. Cela a

comme conséquence l'attribution d'une seule et même étiquette à des unités aussi disparates sur le plan morphologique et flexionnel que lesquels, desquelles, qui, quoi, dont, où, ce qui supprime la possibilité de retrouver les éléments qu'elles pronominalisent en fonction de leurs genre et nombre (pluriel ou invariant).

- L'intégration des informations liées à la personne et au nombre dans les étiquettes relatives aux pronoms et adjectifs possessifs pourrait trouver sa justification en imaginant des requêtes basées sur la recherche de la coréférence/non-coréférence entre un tel pronom ou adjectif et le sujet ou le COI de l'énoncé. Ainsi, nous pourrions trouver par exemple des séquences où le sujet et le pronom possessif partagent les mêmes personne et nombre (Je lui donne le mien) ou non (Je lui donne les vôtres) ou encore des séquences dans lesquelles la coïncidence entre personne et nombre existe entre le COI et l'adjectif possessif (Je lui rends ses clés).

- L'introduction d'étiquettes relatives au partitif du et au déterminant indéfini pluriel des permet à un utilisateur de faire des requêtes basées sur la reconnaissance de constructions partitives ou de syntagmes nominaux indéfinis au pluriel. Il s'avère par ailleurs utile de différencier du partitif et des déterminant indéfini de leurs homographes, les articles définis amalgamés à la préposition de (la sœur du voisin ; la sélection des locataires). Enfin, le codage de la forme faible de l'indéfini pluriel pourrait servir à rechercher les syntagmes nominaux indéfinis où le déterminant précède immédiatement l'adjectif (de belles surprises) sans que celui-ci forme avec le substantif-tête un nom composé (des gros mots). Puisque de et des indéfinis pluriels sont codés de la même façon dans notre corpus, pour obtenir le résultat de la requête imaginée ci-dessus, il faudra restreindre celle-ci en indiquant qu'on cherche tous les éléments ayant l'étiquette DETIIGP et coïncidant avec la chaîne de caractères de. Enfin, l'étiquette DETINEG semble utile pour la recherche de l'élément de qui, dans le cas de la négation absolue, est le substitut de un, du partitif et des indéfini dans les propositions affirmatives correspondantes.

3.2. Limites de ce travail d'enrichissement

Les enrichissements que j'ai effectués dans le cadre de l'annotation morphosyntaxique de notre corpus d'oral transcrit se réduisent à l'ajout d'informations grammaticales et de catégories manquantes basé sur la présentation de l'ensemble des parties du discours dans les grammaires traditionnelles (la grammaire de Riegel et al. a été prise comme source principale). Ayant adopté ce cadre de travail, je n'ai pas proposé l'introduction de nouvelles étiquettes morphosyntaxiques qui rendent compte de l'utilisation particulière que semblent avoir certaines unités à l'oral. Cette décision n'est pas sans poser un certain nombre de problèmes : par exemple, la valeur de l'élément et ne semble pas être la même dans une coordination et en tant qu'élément qui débute un tour de parole ; des mots comme voilà ou quoi dans des suites comme et puis voilà quoi semblent difficilement analysables avec les catégories traditionnelles. Des solutions séduisantes de regroupement de telles unités ont été envisagées, comme par exemple la création d'étiquettes comme « insert » ou « marqueur de relation ». Néanmoins, ces catégories spécifiques ne peuvent pas, à mon sens, rendre compte de la disparité des éléments qu'elles sont censées contenir : si l'on ne sait pas comment traiter l'emploi de quoi achevant une explication et qu'on lui attribue l'étiquette « insert », on est obligé de faire de même, pour garder l'homogénéité, avec des unités ayant clairement le statut d'interjections comme le euh marquant l'hésitation. D'un autre côté, si l'on décide d'associer « marqueur de relation » à tout élément grammatical capable d'instaurer une relation (qui d'ailleurs semble très commode pour une unité

comme « c'est-à-dire », laquelle semble, tout en étant conjonction de subordination, empiéter sur la catégorie des adverbes), on risque de produire une analyse de granularité assez grossière qui ne tienne pas compte des différences fondamentales qui existent entre prépositions et conjonctions, ne serait-ce que du point de vue de la nature des éléments que celles-ci régissent.

L'introduction de nouvelles étiquettes de catégories morphosyntaxiques qui tiennent compte des emplois particuliers qu'ont certaines unités à l'oral pourrait être envisagée à la suite de l'élaboration d'analyses linguistiques décrivant d'une façon approfondie et argumentée ces emplois en tenant compte de leur récurrence. Elle nécessiterait certainement de dédoubler les codages pour enregistrer conjointement la nature et la fonction des unités étiquetées.

4. CONCLUSION

La diversité de la typologie des erreurs générées par un étiqueteur morphosyntaxique, en l'occurrence Cordial Analyseur 7.0, montre que l'utilisation de ce type d'outil informatique ne pourrait s'avérer véritablement fructueuse qu'associée au travail de linguistes qui assurent (i) la validation des analyses et (ii) la révision et l'enrichissement des données linguistiques étiquetées par le système. Le corpus d'oral transcrit constitué à l'UMR Silex/STL, compte tenu de l'attention particulière portée à la précision et à la finesse de l'annotation morphosyntaxique, pourrait efficacement servir en tant que matériel d'entraînement pour des systèmes qui utilisent des méthodes probabilistes pour l'assignation des étiquettes. En fournissant de tels corpus aux systèmes, nous pourrions espérer que ceux-ci se familiariseraient, au moins dans une certaine mesure, avec les constructions et les phénomènes spécifiques liés à la production orale. Outre cette destination liée à l'apprentissage des systèmes d'annotation syntaxique, des corpus comme le nôtre pourraient être intéressants pour des études linguistiques variées et en particulier pour des recherches ciblées basées sur la reconnaissance de la catégorie morphosyntaxique et des informations grammaticales des unités.

REMERCIEMENTS

Je tiens à remercier particulièrement Nathalie Gasiglia (responsable du projet OuRAL à l'UMR Silex/STL) pour ses compétences, sa rigueur, ses remarques et ses corrections minutieuses qui m'ont permis d'améliorer considérablement ce texte. Je remercie également Pierre Corbin et ma directrice de thèse Danièle Van de Velde pour leurs relectures critiques.

NOTES

1. Depuis Janvier 2006, l'UMR Silex a été intégrée dans l'UMR « Savoirs, Textes, Langage », UMR 8163 du CNRS.
2. OUtils et Ressources pour l'Analyse de la Langue. Ce projet s'intègre dans l'action Technolangue. Il est enchâssé dans le projet AGILE coordonné par B. Normier. OuRAL implique six partenaires : Sinequa SA, le Laboratoire de Psychologie Expérimentale (CNRS UMR 8581 – Université Paris V), le Laboratoire d'Informatique de Paris 6 (Université Paris 6), le Laboratoire Informatique d'Avignon, le VALORIA (EA 2593 – Université de Bretagne Sud) et l'UMR Silex/STL.
3. Ces événements correspondent à ceux présentés dans (Gasiglia 2006).
4. N. Gasiglia (responsable du projet OuRAL à l'UMR Silex/STL), O. Cavallaro et moi-même.

5. « Le texte est lu jusqu'à ce qu'une fin de phrase apparaisse, puis il est traité (…). Après constitution de la phrase, le texte est désambiguïsé puis corrigé » (Manuel du logiciel, p. 6).
6. Nous partageons l'idée selon laquelle ponctuer de l'oral transcrit n'est pas pertinent (cf. Blanche-Benveniste et Jeanjean (1987)).
7. À son tour, la désambiguïsation entre les formes homographes du subjonctif et de l'impératif dépend de la présence ou de l'absence d'un pronom personnel devant le verbe.
8. Faire sécher un aliment par la fumée.
9. Il me semble que cette difficulté pourrait être surmontée dans la plupart des cas moyennant la mise en place dans le système de règles syntaxiques distributionnelles comme par exemple : si la forme verbale suivant t' est différente de la 2e personne du singulier, alors lemme = « te » ; sinon, si cette forme est la même pour la 1e et la 2e personne du singulier alors voir si t' est précédé de je ou bien d'un autre élément et en fonction de cette donnée lemme = « te » ou = « tu » ; sinon lemme = « tu ».
10. Dans le cas de (10) donné comme exemple ici, il semble que la désambiguïsation pourrait être assurée par l'affinement des seules règles syntaxiques, à savoir par exemple que si le verbe suivant que n'a pas de complément alors même qu'il s'agit d'un verbe transitif direct, alors que reçoit l'interprétation de relatif (puisqu'il joue lui-même le rôle de complément). Mais cette règle pourrait facilement être violée par l'emploi intransitif (avec ellipse du complément) de verbes transitifs directs (ex : Il est dommage qu'ils cassent). La reconnaissance de tournures impersonnelles comprenant un substantif semble pouvoir contribuer à la levée de ce type d'ambiguïté.
11. Ces ajouts ont été validés durant nos séances de travail par N. Gasiglia et O. Cavallaro.
12. L'étiquette originale de CA7.0 est DETPOSS et non DTPS. Nous avons décidé de la condenser afin de ne pas obtenir des étiquettes trop longues après les ajouts effectués.

RÉFÉRENCES BIBLIOGRAPHIQUES :

Benzitoun C. 2004. « L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? », *actes de la conférence RECITAL*, Fès.

Benzitoun C. et al. 2004. « L'analyse syntaxique de l'oral : problèmes et méthode », *Journée ATALA du 15 mai 2004*.

Blanche-Benveniste C. et Jeanjean C. 1987. *Le français parlé. Transcription et édition*, Paris : Didier érudition.

Clement L. 2001. *Construction et exploitation d'un corpus syntaxiquement annoté pour le français*, thèse présentée à l'Université Paris VII pour obtenir le titre de docteur, spécialité linguistique.

Fuchs C. 1996. *Les ambiguïtés du français*, Paris : Ophrys.

Gasiglia N. 2006. « Stratégie de consultation de corpus oraux transcrits : pistes méthodologiques pour l'exploration d'un corpus thématique à haut rendement », dans Williams G. (ed.) 2006. *Texte et Corpus*, Rennes : Presses Universitaires de Rennes.

Habert, B. 2005. *Instruments et ressources électroniques pour le français*, Paris : Ophrys.

Riegel M., Pellat J-C. et Rioul R. 1994. *Grammaire méthodique du français*, Paris : PUF.

Wagner R., Pinchon J. 2002. *Grammaire du français classique et moderne*, Paris : Hachette.

Le Grand Robert de la langue française, 1994 (version électronique), Paris : Le Robert**.**

Le Petit Robert, 1996 (version électronique), Paris : Le Robert.

Dictionnaire de la langue française (LEXIS), 1979, Paris : Larousse.