TEXTE ET CORPUS, N°4

Actes des sixièmes Journées de la Linguistique de Corpus

Sous la direction de Geoffrey Williams

2013

Direction : Geoffrey Williams
Responsable de l'Édition : Araceli Alonso Campo

Janvier 2013

Maison de la Recherche (Bâtiment Paquebot)
Université de Bretagne-Sud
4, rue Jean Zay
56321 Lorient Cedex (France)

ISSN : 1958-5306

SOMMAIRE

ACCÉDER À UN CORPUS DE TEXTES SCIENTIFIQUES PAR UNE RESSOURCE TERMINO-ONTOLOGIQUE MULTI-PLANS

Vanessa Andréani
Laboratoire LIDILEM (EA 609) – Université Stendhal Grenoble 3
TecKnowMetrix SAS – Voiron

RÉSUMÉ

Un nombre important de structures, industrielles ou non, doivent gérer et exploiter de grands ensembles documentaires. Ces ensembles peuvent être considérés comme des anagnoses, puisqu'ils recensent l'historique textuel de ces structures. Leur exploitation peut être facilitée par la modélisation des documents et la représentation des connaissances qu'ils contiennent, à travers la construction d'une ressource termino-ontologique (RTO). Nous proposons une RTO multi-plans permettant d'accéder à l'information pertinente, tout en conservant les spécificités des données qu'elle contient. L'utilisateur peut ainsi accéder à l'information en fonction d'un point de vue donné, adapté à son besoin immédiat. L'utilisation finale de cette RTO consistera en une visualisation dynamique des informations récupérées par l'utilisateur grâce à la représentation en plans des données.

1 INTRODUCTION

Un nombre important de structures, industrielles ou non, doivent gérer de grands ensembles documentaires. Si leur stockage est de moins en moins problématique grâce à des capacités de stockage toujours plus importantes, leur maintenance et leur exploitation soulèvent des difficultés, notamment en raison de la quantité de données à parcourir pour obtenir l'information pertinente sur une question précise.

Pour accéder à l'information textuelle contenue dans un ensemble documentaire, nous proposons de modéliser les documents en fonction d'un ou de plusieurs points de vue. Cette modélisation peut être faite par la construction d'une ressource termino-ontologique (RTO), c'est-à-dire une ressource permettant de représenter de manière organisée un certain nombre de concepts et de termes associés concernant un domaine. Cette ressource regroupe les éléments pertinents pour la recherche et la consultation de documents.

Nous avons opté pour une RTO multi-plans qui permet d'accéder à l'information adéquate à un besoin donné. Les éléments pertinents sont extraits du corps du texte, ou directement des données associées.

Notre RTO est structurée en plusieurs plans en fonction du type de données, de manière à fournir un accès adapté à l'utilisateur quelle que soit sa demande.

Dans une première partie de cet article, nous présenterons l'ensemble documentaire de la société TecKnowMetrix, qui représente notre corpus de travail. Dans un deuxième temps, nous détaillerons la structure de notre RTO multi-plans, en appuyant sur l'intérêt d'une représentation en plusieurs plans. Enfin, nous présenterons nos conclusions à ce stade et nos perspectives de travaux pour les mois à venir.

2 L'ENSEMBLE DOCUMENTAIRE COMME CORPUS DE TRAVAIL

TecKnowMetrix (TKM) est une jeune entreprise innovante qui propose à ses clients des prestations de conseil en stratégie de l'innovation. Ces prestations vont de l'état de l'art sur des domaines innovants à des études de marché en passant par des analyses concurrentielles. Le point commun à ces différentes prestations est l'analyse de grands corpus de brevets, articles scientifiques et articles de presse technico-économique, comportant le plus souvent plusieurs milliers de textes. Ces corpus sont constitués par ses analystes afin de regrouper de façon exhaustive la production scientifique et technique gravitant autour des domaines d'intérêt de ses clients.

TKM a constitué dans le cadre de ses études plusieurs centaines de corpus regroupant plusieurs millions de documents. L'ensemble de ces corpus forme en quelque sorte le « passé textuel » de la société, une forme d'*anagnose* au sens de (Thlivitis, 1998). Une anagnose est vue comme le regroupement de l'ensemble des textes parcourus par un individu. Elle oriente très fortement la lecture de nouveaux textes puisqu'elle guide les parcours interprétatifs au sens de (Rastier, 2001, pages 110-111) en permettant certaines actualisations sémantiques. Prendre en considération l'anagnose, ou au moins une approximation de celle-ci, se révèle ainsi primordiale à nos yeux dans le cadre d'analyses approfondies de corpus.

L'ensemble documentaire scientifique et technique regroupant les corpus associés aux études TKM constitue notre corpus de travail. Cet ensemble documentaire n'est pas indexé : toutes les recherches sont effectuées en texte plein. En effet, les utilisateurs experts ont besoin de faire des recherches extrêmement fines, pour lesquelles la position des mots dans le texte et leur ordre doit être conservé. De plus, il arrive fréquemment qu'un terme recherché soit composé de huit mots et plus, comme c'est le cas pour un composé chimique. De fait, une indexation serait peu utile pour ce type de recherches, et pourrait même créer des parasites dans les résultats.

Les documents sont agencés au sein d'une base de données, et chaque corps de texte est associé *via* différentes tables à des informations telles que son auteur, l'organisation de laquelle il émane, sa date et son pays de publication, etc., que nous définissons comme les méta-données associées à chaque texte. C'est à partir de ce corpus, regroupant le passé textuel de TKM, que nous proposons une approximation de l'anagnose de la société à travers une ressource termino-ontologique (RTO). Les RTO modélisent à la fois les concepts d'un domaine, comme pour une ontologie, et les termes qui les représentent, ainsi que le ferait une terminologie (Tissaoui, 2009 ; Aussenac-Gilles *et al.*, 2006). Les RTO regroupent en réalité des terminologies aux structures plus ou moins complexes, tels des réseaux lexicaux, des bases de données lexicales et sémantiques, etc.

Notre travail porte sur les données de la société TKM, mais pourrait être adapté à d'autres collections de documents : des bibliothèques numériques, de grandes collections de pages Web, de blogs, etc.

3 UNE RESSOURCE TERMINO-ONTOLOGIQUE MULTI-PLANS

Cette RTO a été constituée afin de permettre aux utilisateurs experts de gérer et de consulter l'ensemble documentaire de manière intuitive, simple et adaptée à leurs besoins. De fait, notre objectif est de tirer de cette RTO et de sa projection en corpus des informations permettant de :

1. capitaliser l'information antérieure, de manière à guider et orienter les analyses et interprétations des utilisateurs ;
2. fournir aux utilisateurs un accès multiple, simplifié et plus intuitif à ces documents, notamment à l'aide de projections cartographiques comme proposé dans (Roy, 2007).

D'un point de vue technique, cette RTO est structurée en base de données MySQL. Les informations de chacun des plans que nous allons présenter sont stockées dans une table dédiée. Un identifiant représentant le document est associé à chaque information extraite de celui-ci, et ce quel que soit le plan. Dans notre corpus, un document contient forcément au moins une information de chaque plan. De fait, chaque plan est relié à la fois à l'ensemble documentaire et aux autres plans, par le biais des documents et de leurs identifiants.

3.1 Un plan pour chaque type de données

Notre RTO est constituée de cinq plans liés entre eux, et se présente sous la forme d'une base de données. Chacun des plans est construit à partir d'un type d'information tiré de notre corpus de travail, soit à partir du corps des documents, soit dans ce que nous avons appelé précédemment les méta-données associées. Chaque plan regroupe des entités de même nature qui constituent selon nous des points d'entrée pertinents pour l'analyse du corpus. Les cinq plans sont présentés ci-dessous.

3.1.1 LES DONNÉES INTERNES AUX DOCUMENTS : LE PLAN THÉMATIQUE

Le plan thématique recense les thèmes de chaque texte. Comme dans (Pichon et Sébillot, 1999), nous considérons le thème d'un texte comme l'ensemble des sujets qui y sont abordés. Ces thèmes sont représentés par les segments répétés issus de chaque texte. En effet, nous partons du principe qu'un terme qui est présent plus d'une fois dans le même document est pertinent en tant qu'il représente une partie du thème abordé par le texte.

Afin d'éviter de ramener trop de données bruitées, et à la différence de l'inventaire des segments répétés de (Lafon et Salem, 1983), nous avons exclu des segments répétés les suites commençant ou finissant par un mot vide au sens de (Tesnière, 1969). Un segment répété peut contenir de 2 à 9 tokens, et doit être présent au moins deux fois dans un même document pour être considéré comme tel. A la suite de (Lafon et Salem, 1983) cette fois, nous excluons les segments contenant des signes de ponctuation. La limite de 9 tokens autorisés a été définie suite à des observations en corpus : lorsque nous rencontrons des tokens très longs, ils ne dépassent généralement pas 8 ou 9 tokens. C'est notamment le cas des composés chimiques, fréquemment cités dans notre ensemble documentaire scientifique et technique.

En pratique, chaque segment répété est associé dans la base de données au(x) document(s) duquel / desquels il est issu.

3.1.2 LES MÉTA-DONNÉES ASSOCIÉES AUX DOCUMENTS

Le plan des organisations

L'organisation de laquelle émane le document est identifiée en amont dans une table dédiée et est reliée à « son » document. Une organisation peut être une entreprise ou une structure institutionnelle, telle une université ou un hôpital. Il peut également s'agir d'un particulier, dans le cas où cet individu n'est rattaché à aucune structure collective et publie un article ou dépose un brevet en son nom propre. Dans tous les cas, nous traitons donc des entités nommées (EN), qui sont, d'après (Poibeau, 2001), « *l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné* ». L'auteur inclut également dans cette catégorie « *les dates, les unités monétaires, les pourcentages, etc.* ». Tous les noms

d'organisations sont normalisés (Andréani, 2009), ce qui permet d'obtenir des données fiables et nettoyées, puis sont intégrés au plan des organisations.

Ce plan permet de mettre en exergue les interactions entre plusieurs organisations, ce qui est un besoin récurrent des utilisateurs de la base, par exemple pour établir des réseaux de collaboration.

Le plan des auteurs

Sur le même principe que pour les organisations, les auteurs sont stockés dans une table dédiée liée aux documents. D'un point de vue formel, il s'agit là encore d'EN, de personnes cette fois. Du point de vue conceptuel, la différence principale entre une organisation et un auteur réside dans le fait que de manière générale, un auteur est un individu rattaché à une organisation dans le cadre de la publication d'un document. Notons qu'un même individu peut être rattaché, simultanément ou non, à plusieurs organisations, et qu'une organisation peut compter un grand nombre d'auteurs. Il est donc important de pouvoir faire le lien entre une organisation et ses auteurs, de même que de pouvoir mettre en valeur les interactions entre auteurs. Là encore, les noms d'auteurs sont extraits et stockés dans l'ontologie.

Le plan géographique

Les informations géographiques concernant un document sont en réalité les coordonnées de l'organisation qui l'a publié. Nous exploitons donc la table des organisations, qui contient, outre leur nom, des renseignements tels que leur adresse. Nous extrayons ces données, et en particulier la ville et le pays d'origine de l'organisation, soit des EN de lieu. Ces données ne sont pas présentes sous forme standardisée, et une première phase de travail consiste donc à repérer et normaliser ces informations. Dans un second temps, l'identifiant du document est associé à un point de l'espace dans la RTO géographique, qui est le seul plan à être constitué *a priori*, à partir de listes de villes et pays. Chaque jeu de coordonnées ville – pays est associé à ses coordonnées en longitude et latitude, dans le but de pouvoir projeter les données relatives à des documents sur une carte géographique.

Le plan temporel

Les informations concernant la date de publication des documents se trouvent dans un champ dédié dans la table contenant le corps des documents, et se présentent sous une forme standardisée. Elles sont extraites et répertoriées, toujours en lien avec le document courant, dans la RTO temporelle. Ce plan sert avant tout de filtre pour la visualisation des données de la RTO dans sa totalité. Par exemple, l'utilisateur pourra s'intéresser aux publications d'un sous-ensemble d'organisations donné sur une période réduite de 10 ans.

3.2 Combiner les plans pour révéler l'information

L'objectif de cette RTO est de fournir un accès multiple et intuitif aux documents de la base. La structuration en plans permet d'accéder aux documents par un point de vue déterminé en fonction du besoin immédiat de l'utilisateur expert.

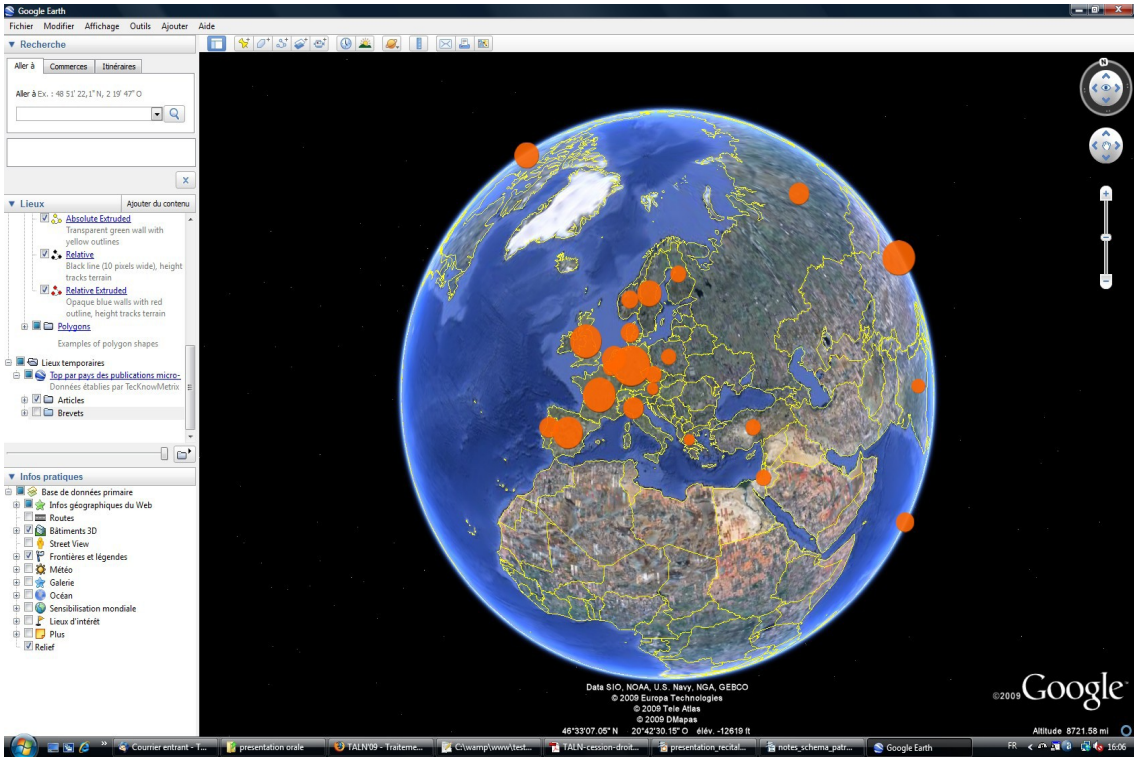
Ainsi que nous l'avons mentionné, si l'utilisateur a besoin de repérer les réseaux de collaboration entre organisations, il peut accéder à cette information par le plan des organisations. Les documents se situant à l'intersection de plusieurs entreprises ou institutions publiques sont l'indice de ces collaborations. De même, il peut savoir quels auteurs travaillent ensemble en utilisant le plan des auteurs individuels.

Le fait que tous les plans soient liés entre eux ouvre également un grand nombre de possibilités de combinaisons dès lors que l'utilisateur en éprouve le besoin. Il peut ainsi accéder à une information qu'il lui aurait été difficile d'obtenir par la consultation linéaire des documents ou par la création de tableaux. Nous donnons ici quelques exemples de

combinaisons, qui ne représentent pas une liste exhaustive des possibilités offertes par la RTO.

1. La combinaison du plan des organisations avec celui des auteurs peut permettre à l'expert de mettre au jour des liens entre certaines organisations lorsqu'un auteur a écrit un article ou déposé un brevet au nom de plusieurs organisations. Ces liens peuvent être un point de départ si le but de l'utilisateur est de déceler des possibilités de collaborations entre entreprises.
2. De même, pour déterminer les auteurs qui interviennent à la fois dans des structures institutionnelles et des entreprises autour d'une même thématique, il est possible de combiner le plan des auteurs, celui des organisations et celui des thèmes des documents. De cette manière, tous les auteurs liés à des documents contenant les mêmes termes et travaillant dans plusieurs organisations seront repérés.
3. Enfin, il est possible de localiser géographiquement les organisations les plus actives pour un thème donné, et ainsi de dégager, à partir du plan géographique et de celui des organisations, les pays ou les régions du monde les plus performants dans un domaine déterminé.

Par la suite, ces combinaisons pourront être visualisées grâce à des cartographies géographiques ou sémantiques, dont nous présentons deux exemples en figure 1. Ces visualisations seront dynamiques, et permettront de faire émerger des informations pertinentes qui n'auraient pas été décelées par une recherche d'information « classique ». A partir de cette première visualisation, l'utilisateur peut accéder aux documents concernés.



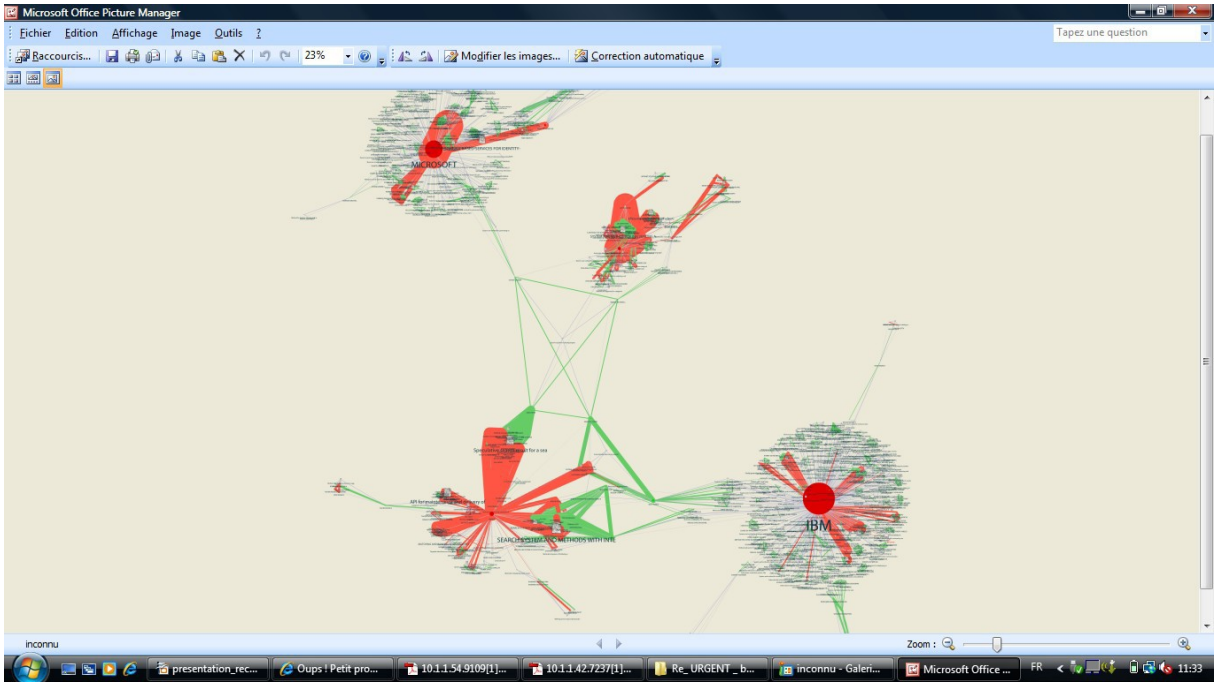


Figure 1: Répartition par pays des travaux sur les biocarburants (en haut) et réseau de collaboration entre IBM et Microsoft (en bas)

Cette RTO s'applique en pratique à l'ensemble documentaire de la société TKM. Cependant, son principe et sa structure peuvent être reproduits dans d'autres cas, dès lors que l'utilisateur a besoin de prendre en compte plusieurs types d'informations et de les recouper.

4 CONCLUSION ET PERSPECTIVES

L'ensemble documentaire constitué par les analystes de TKM au fil de leurs différentes études constitue une sorte d'anagnose de la société. A ce titre, elle représente le savoir et les connaissances accumulés depuis la création de l'entreprise, et de ce fait influence l'interprétation de nouveaux documents.

Il est primordial pour TKM de gérer au mieux cette anagnose et de pouvoir en tirer des informations pertinentes et adaptées aux besoins immédiats des utilisateurs experts. Cet accès aux documents et à l'information qu'ils véhiculent peut être grandement facilité par une représentation formelle de ces documents, à travers une ressource termino-ontologique (RTO) multi-plans. Chacun des plans de cette RTO contient un type déterminé de données. De cette façon, l'utilisateur peut accéder aux documents d'un point de vue pertinent pour l'objectif fixé par les besoins d'une étude précise.

A court et moyen terme, nos travaux porteront sur la conception d'un outil complet de visualisation qui s'appuiera sur la RTO, à la fois à partir des informations elles-mêmes, mais aussi de la manière dont ces informations sont structurées.

L'accès aux informations pertinentes et au résultat de leur combinaison se fera par le biais d'une visualisation cartographique dynamique à partir des critères rentrés par l'utilisateur, et pourra prendre plusieurs formes en fonction de ses besoins. Par exemple, des documents traitant un sujet précis seront projetés sur une carte géographique selon les lieux dont ils sont issus, ce qui permettra à l'utilisateur de déterminer les régions du monde les plus performantes dans un domaine donné, qu'il aura lui-même déterminé. Dans ce cas, le plan géographique sera couplé au plan thématique. L'outil final sélectionnera automatiquement le type de cartographie le plus adapté aux types de données sélectionnés, tout en laissant la possibilité à l'utilisateur de choisir lui-même une visualisation précise. L'un des aspects

fondamentaux de cet outil résidera dans son caractère dynamique et interactif : les visualisations seront générées à la demande, à partir d'une requête, puis permettront d'accéder au(x) document(s) sélectionné(s) en cliquant sur la cartographie.

Rappelons que le principe et la structure de notre RTO multi-plans peuvent être appliqués à d'autres ensembles documentaires, à partir du moment où les types de données pertinents pour une recherche ont été identifiés et extraits en amont. Par exemple, nous avons pour projet à moyen terme de l'adapter dans le cadre du projet des Manuscrits de Stendhal en ligne, qui rendra accessible les manuscrits de l'auteur au grand public et aux spécialistes de la littérature. Si les entités pertinentes pour cet ensemble documentaire diffèrent de celles de TKM, leur structuration en plans n'en demeure pas moins adaptée aux besoins des utilisateurs.

Le fait de structurer notre modèle de représentation des connaissances en différents plans permet donc à la fois d'avoir une ressource cohérente à considérer comme une unité, et de restituer la diversité des éléments qui y sont stockés en fonction de leurs spécificités.

5 RÉFÉRENCES

Andréani V. (2009). « Normalisation des entités nommées : pour une approche mixte et orientée utilisateurs ». *RECITAL 2009, Actes de TALN 2009*. France : Senlis.

Aussenac-Gilles N., Condamines A. et Sèdes, F. (2006). « Évolution et maintenance des ressources termino-ontologiques : une question à approfondir ». *Information - Interaction - Intelligence, Numéro spécial Ressources termino-ontologiques.* Hors-série. p. 7-14.

Lafon P. et Salem A. (1983). « L'Inventaire des segments répétés d'un texte ». *Mots,* N°6, p. 161-177.

Pichon R. et Sébillot P. (1999). « Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience ». *Actes de TALN 1999*. p. 279-288.

Poibeau T. (2001). « Deconstructing Harry, une évaluation des systèmes de repérage d'entités nommées ». *Revue de la société d'électronique, d'électricité et de traitement de l'information*.

Rastier F. (2001). *Arts et sciences du texte*. Paris : Presses Universitaires de France.

Roy T. (2007). Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires. Thèse de Doctorat, Université de Caen Basse-Normandie.

Tesnière L. (1969). *Éléments de syntaxe structurale.* Paris : Klincksieck.

Thlivitis T. (1998). *Sémantique interprétative Intertextuelle : Assistance anthropocentrée à la compréhension des textes.* Thèse de Doctorat en Informatique, Université de Rennes I, Rennes.

Tissaoui A. (2009). « Typologie de changements et leurs effets sur l'évolution de Ressources Termino-Ontologiques ». Dans F. Gandon (éds.), *IC 2009 : Posters des 20es Journées Francophones d'Ingénierie des Connaissances*, *Hammamet (Tunisie)*.

ENRICHIR LES DESCRIPTIONS LEXICOGRAPHIQUES À PARTIR DE CORPUS PARALLÈLES ET CIBLÉS : *AUGMENTER* & CO.

Ann Bertels[1], Cédrick Fairon[2], Jörg Tiedemann[3] et Serge Verlinde[4]
1. ILT et QLVL – K.U.Leuven (Belgique)
2. CENTAL – UCL Louvain-la-Neuve (Belgique)
3. Department of Linguistics and Philology – Uppsala University (Suède)
4. ILT – K.U.Leuven (Belgique)

RÉSUMÉ

L'objectif de cet article est de présenter quelques-unes des nouvelles techniques d'analyse et d'exploitation de corpus, dans le but d'enrichir les descriptions lexicographiques traditionnelles des verbes dénotant la notion de hausse ('trends verbs'), pour l'anglais, le français et le néerlandais. A cet effet, nous procédons à une analyse de corpus parallèles et de corpus monolingues ciblés. Les corpus parallèles fournissent des indications sur la catégorie grammaticale des traductions des verbes dénotant la notion de hausse, ainsi que sur la fréquence d'emploi et l'équivalence de ces traductions. Les corpus monolingues ciblés permettent de relever les collocatifs significatifs et nous renseignent dès lors sur la combinatoire et les contextes d'emploi des verbes de hausse. Les résultats des différentes analyses de corpus, en termes de profils de traduction et profils combinatoires, contiennent des indications précieuses pour enrichir les descriptions lexicographiques traditionnelles dans les dictionnaires de traduction.

1 INTRODUCTION

Les analyses de corpus ont révolutionné la linguistique et la lexicographie (Cf. Rundell, 1998 ; Béjoint, 2004). A la fin des années 1970, les dictionnaires COBUILD ont commencé à appuyer leurs descriptions lexicographiques sur l'analyse de corpus textuels de quelques millions de mots. Depuis les années 1980-1990, on a vu se généraliser une approche lexicographique basée sur des analyses de corpus ('corpus-driven' et 'corpus-based lexicography'). Les approches 'corpus-based' fournissaient des exemples authentiques et les approches 'corpus-driven' permettaient même de découvrir des régularités ou patrons linguistiques.

Toutefois, ces dernières années, les techniques de collecte et d'analyse de données ont beaucoup évolué. Un nombre toujours grandissant de documents numériques sont disponibles et accessibles sur Internet. Pensons notamment au British National Corpus (BNC)[1], à Frantext[2], à Webcorp ou aux initiatives de Web utilisé comme corpus ('Web as corpus') (Kilgarriff et Grefenstette, 2003 ; Hundt, Nesselhauf et Biewer, 2007). Les corpus électroniques ne demandent qu'à être exploités et analysés, ce qui est facilité considérablement par la mise à disposition du grand public d'outils d'analyse très performants,

[1] http://www.natcorp.ox.ac.uk
[2] http://atilf.atilf.fr/frantext.htm

tels que Lexico3[3], Hyperbase[4], WordSmith[5] ou même le logiciel d'analyse statistique Open Source R[6].

Dans cet article, nous décrivons une étude exploratoire dans le domaine de la lexicographie bilingue (Cf. Bertels, Fairon, Tiedemann et Verlinde, 2009). L'objectif est de présenter quelques-unes des nouvelles techniques d'analyse et d'exploitation de corpus, qui permettent d'enrichir les descriptions lexicographiques traditionnelles des verbes dénotant la notion de hausse ('trends verbs') (Verlinde, 1995), pour l'anglais (*increase*, *rise*, etc.), le français (*augmenter*, *progresser*, etc.) et le néerlandais (*verhogen*, *stijgen*, etc.). D'abord, nous relevons les principales lacunes des dictionnaires de traduction traditionnels (section 2). Ensuite, nous présentons les corpus utilisés dans cette étude exploratoire, à savoir des corpus parallèles et des corpus monolingues ciblés (section 3). Finalement, nous discutons les résultats des analyses de corpus en termes de profils de traduction et profils combinatoires (section 4). Nous terminons par la conclusion (section 5).

2 LES LACUNES DES DICTIONNAIRES DE TRADUCTION

Le point de départ de notre expérimentation est la constatation de quelques lacunes dans les dictionnaires de traduction traditionnels, principalement dans le domaine des traductions et des contextes d'emploi. Ces lacunes se retrouvent dans plusieurs dictionnaires de traduction que nous avons consultés, notamment Oxford-Hachette, Robert & Collins, Robert & Van Dale et Van Dale vertaalwoordenboeken.

Premièrement, en ce qui concerne les traductions, les dictionnaires de traduction donnent généralement un ou plusieurs mots équivalents dans la langue cible (L2). Ces équivalents, qui relèvent quasi toujours de la même catégorie grammaticale, sont parfois regroupés par contexte ou en fonction d'une indication de sens (Cf. figure 1).

Figure 1 : Extrait du Van Dale (français-néerlandais) pour le verbe progresser

La figure 1 montre, à titre d'exemple, les traductions du verbe français *progresser* en néerlandais. Le deuxième sens, qui est un sens figuré, se traduit d'abord par un verbe indiqué en gras, ensuite par deux autres verbes et un substantif (*vorderingen*). Un apprenant de FLE (français langue étrangère) aura donc tendance à traduire le verbe *progresser* par *vooruitgaan*, puisqu'il est mis en évidence par la typographie du dictionnaire. Or, les analyses de corpus parallèles font ressortir l'importance d'autres traductions.

Les descriptions traditionnelles des dictionnaires de traduction présentent plusieurs lacunes en ce qui concerne les traductions proposées. En effet, l'utilisateur n'est pas informé sur la fréquence d'emploi des traductions ni sur l'équivalence des traductions, ce qui s'avère surtout

[3] http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/
[4] http://ancilla.unice.fr/~brunet/pub/hyperbase.html
[5] http://www.lexically.net/wordsmith
[6] http://www.r-project.org

problématique pour des mots polysémiques. Par ailleurs, l'utilisateur n'est pas informé non plus sur la pertinence de traductions appartenant à d'autres catégories grammaticales, si celles-ci sont mentionnées. Généralement, les verbes sont traduits par des verbes, comme le montre l'exemple du verbe français *progresser* (Cf. figure 1). Finalement, on observe un manque de cohérence et de systématicité, puisque d'un dictionnaire à l'autre, les traductions varient sensiblement. Le recours à des analyses de corpus parallèles pourrait nous aider à combler ces lacunes, par exemple pour les verbes dénotant la notion de hausse. Les corpus parallèles permettent d'étudier les traductions de ces verbes de hausse et dès lors de déterminer leurs profils de traduction.

Deuxièmement, en ce qui concerne les contextes d'emploi, les dictionnaires de traduction essaient déjà de contextualiser les traductions (Cf. figure 1). Toutefois, ces contextes sont souvent donnés en vrac, sans différenciation sémantique, ni phrases-exemples (Cf. figure 2 pour le verbe *augmenter*). Il serait intéressant d'enrichir ces informations contextuelles à l'aide de profils combinatoires plus précis, basés sur les collocatifs statistiquement significatifs. Ces profils combinatoires pourraient être établis à partir d'analyses de fréquence et d'analyses statistiques (exploratoires) de corpus monolingues ciblés.

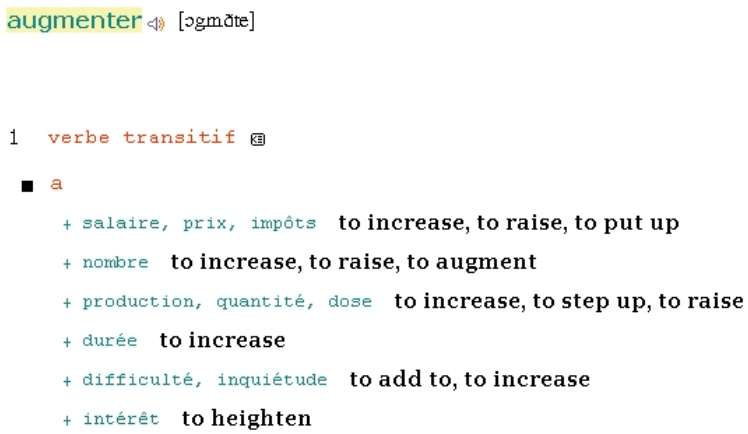


Figure 2 : Extrait du Robert & Collins (version CD-ROM)

3 LES CORPUS PARALLÈLES ET CIBLÉS

3.1 Les corpus parallèles

Les corpus parallèles que nous avons analysés pour établir les profils de traduction appartiennent au projet OPUS[7] (Tiedemann et Nygaard, 2004), qui a pour but de rassembler des corpus parallèles librement accessibles et de les enrichir d'outils de recherche. Le projet OPUS regroupe plusieurs textes multilingues 'Open Source', notamment des textes informatiques, juridiques et biomédicaux et des sous-titres. Dans cette expérimentation par rapport aux verbes de hausse, nous avons analysé des comptes rendus du parlement européen (Europarl). Les textes sont fournis dans un format XML standardisé. OPUS se caractérise par une approche novatrice qui réside dans un alignement au niveau du mot. Les liens de mot à mot ont été obtenus à l'aide de l'outil Giza++[8], utilisé dans le domaine de la traduction automatique statistique.

[7] http://urd.let.rug.nl/tiedeman/OPUS
[8] http://code.google.com/p/giza-pp

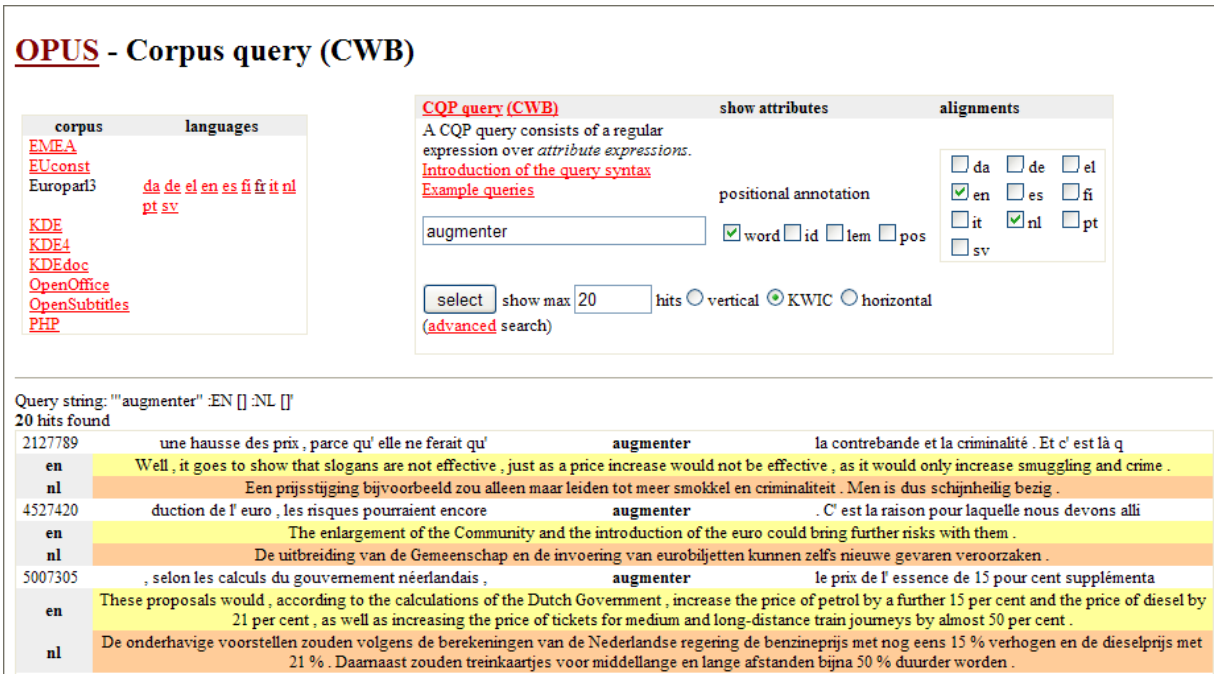


Figure 3 : OPUS : visualisation d'une requête dans les corpus parallèles Europarl

La figure 3 montre une capture d'écran d'une requête réalisée dans OPUS dans les corpus parallèles Europarl, pour le verbe *augmenter* et ses traductions en anglais et en néerlandais. Les phrases-exemples authentiques des corpus parallèles montrent que le verbe *augmenter* à l'infinitif ne se traduit pas uniquement par des formes verbales (Cf. *increase* dans la première phrase-exemple), mais aussi par des adjectifs (Cf. *meer* dans la première phrase et *further* dans la deuxième phrase-exemple). Toutes les phrases-exemples contenant un verbe de hausse (en français, en anglais et en néerlandais) ont été rassemblées dans une base de données, avec les traductions dans les deux autres langues. Cette base de données est interrogeable via une interface web et des scripts en php. Ainsi, par lemme dans la langue-source, on peut extraire les traductions dans les deux langues-cibles, grâce à l'alignement au niveau du mot. Ces requêtes permettent de générer des listes de paires de mots (verbe-traduction), classées par fréquence décroissante de cooccurrence.

3.2 Les corpus monolingues ciblés

Les corpus monolingues ciblés proviennent du Web et ils ont été rassemblés par le logiciel Corporator (Fairon, 2006), qui effectue une analyse d'un certain nombre de sources prédéfinies : des flux RSS. Les corpus utilisés dans le cadre de cette expérimentation relèvent du domaine économique et se constituent de flux RSS en anglais (*The Economist*), en néerlandais (*BNR*, *de Standaard*) et en français (*Le Monde*, *La Libre Belgique*, *La Tribune*). Nous avons ciblé des textes du domaine économique, parce qu'ils se caractérisent par un emploi important de verbes de hausse. Nous avons collecté des articles journalistiques entre mai 2007 et mai 2008, pour avoir un corpus d'une certaine taille. Ensuite, le logiciel 'Open Source' Unitex (Paumier, 2003) a été utilisé pour effectuer des recherches et construire des fichiers de concordance, c'est-à-dire pour isoler des phrases contenant des verbes dénotant la notion de hausse.

4 DISCUSSION DES RÉSULTATS

Dans le cadre de l'expérimentation que nous présentons ici, nous avons étudié la traduction de la série de verbes dénotant la notion de hausse suivante : en anglais *increase, raise, rise, boost, recover*, en français *augmenter, progresser, gagner, croître, accroître, agrandir, grandir* et en néerlandais *stijgen*, *toenemen*, *verhogen* et *vergroten*. Les verbes de hausse sont certes plus nombreux, mais nous nous sommes limités aux verbes les plus fréquents qui figurent aussi bien dans les corpus parallèles et que dans les corpus ciblés.

4.1 Les profils de traduction

Afin de déterminer les profils de traduction des verbes de hausse, nous avons eu recours à des corpus parallèles. Dans un premier temps, nous avons analysé les listes de fréquence des paires de mots fournies pour chaque verbe séparément. Cette première analyse plutôt basique a tout de même permis de constater que de 9 à 55% des traductions des verbes de hausse se font à l'aide d'un mot qui appartient à une autre catégorie grammaticale, à savoir des noms, des adjectifs et des adverbes. En effet, le verbe *progresser* se traduit en néerlandais moins souvent par un verbe (45% des traductions) (par exemple *vorderen*) que par un mot appartenant à une autre catégorie grammaticale (55% des traductions). Si tel est le cas, le verbe *progresser* se traduit dans les trois quarts des cas par un nom (par exemple *vooruitgang, vorderingen*) et par un adverbe dans le quart restant (notamment *voort*, *vooruit*, *verder*). Par contre, *augmenter* se traduit dans la majorité des cas par un verbe en néerlandais (84%) (par exemple *verhogen*, *toenemen*, *stijgen*). Lorsqu'il n'est pas traduit par un verbe, il l'est le plus souvent par un adjectif (57%) (*meer*, *groter*) et par un nom (39%) (*verhoging*). Les dictionnaires de traduction devraient non seulement mentionner cette variété de traductions, mais également, dans une version électronique, prévoir des liens vers des phrases-exemples et des contextes d'emploi authentiques, tirés par exemple d'un corpus parallèle disponible en ligne.

La fréquence d'emploi et l'équivalence des traductions constituent également des données intéressantes pour le lexicographe. En effet, comme nous l'avons mentionné ci-dessus, les traductions fournies dans les dictionnaires de traduction traditionnels n'ont pas toutes la même fréquence. Par ailleurs, elles ne sont pas forcément équivalentes dans les différents contextes d'emploi. La question qui se pose donc est celle de savoir comment les verbes de hausse se situent les uns par rapport aux autres. Se caractérisent-ils par des traductions similaires ? Quel est leur profil de traduction ? Compte tenu du nombre de traductions, de leur fréquence très variée et du nombre de verbes dénotant la notion de hausse, il est impossible de concevoir comment ces verbes se comportent les uns par rapport aux autres, donc dans quelle mesure ils présentent des profils de traduction similaires.

La solution réside dès lors dans le recours aux analyses MDS (*MultiDimensional Scaling*). Ces analyses statistiques multidimensionnelles permettent de comparer des relevés de données (ici : la fréquence des traductions) pour plusieurs variables (ici : les verbes de hausse). Le but des analyses MDS est de déterminer dans quelle mesure les variables ont un comportement analogue ou non (ici : dans quelle mesure les verbes ont un profil de traduction semblable ou non). L'analyse MDS est implémentée dans le logiciel d'analyse statistique Open Source R. Le logiciel rend les similarités et les dissimilarités entre les distributions des fréquences de traduction au moyen de petites et grandes distances, qui peuvent être projetées dans un espace à deux dimensions. L'analyse MDS génère donc une matrice de dissimilarité à partir d'une table de contingence, comme celle qui est représentée ci-dessous et qui fait

l'inventaire des fréquences des paires 'verbes français'-'traduction en néerlandais' dans un corpus parallèle aligné au niveau du mot (Cf. tableau 1). Le tableau 1 ne montre qu'un extrait de la table de contingence des verbes français, qui compte au total 48 traductions différentes (en rangées).

	augmenter	progresser	gagner	croître	accroître	agrandir	grandir
stijgen	410	1	1	30	1	1	1
stijging	11	1	1	1	1	1	1
toename	18	1	1	1	1	1	1
toenemen	550	10	1	143	175	1	1
trekken	16	1	1	1	1	1	1
verbeteren	42	1	1	1	35	1	1
verder	1	76	1	1	19	1	1
verdienen	1	1	133	1	1	1	1
vergroten	270	1	1	1	291	1	1
vergroting	20	1	1	1	1	1	1
verhogen	775	1	1	1	187	1	1
verhoging	126	1	1	1	32	1	1

Tableau 1 : Extrait de la table de contingence des verbes français (avec leurs traductions en néerlandais en rangées)

Les rapports entre les verbes analysés, en termes de similarités et dissimilarités, sont ensuite visualisés selon deux axes, qui présentent les données analysées et leurs distances. Les verbes qui se traduisent souvent par les mêmes mots à des fréquences similaires se regroupent en clusters. Par contre, les verbes qui présentent un profil de traduction très spécifique, c'est-à-dire qui se traduisent par des mots différents ou par les mêmes mots mais à des fréquences très différentes, occupent une position isolée ou périphérique par rapport aux autres verbes analysés. La visualisation des rapports entre les verbes français de hausse, en fonction de leurs traductions en néerlandais, montre clairement que trois verbes se distinguent des autres verbes par leur profil de traduction particulier, à savoir *augmenter*, *gagner* et *progresser* (Cf. figure 4). Les traductions des autres verbes de hausse (*grandir*, *croître*, *agrandir*, *accroître*) sont moins exclusives et, par conséquent, ces verbes se situent plus près les uns des autres.

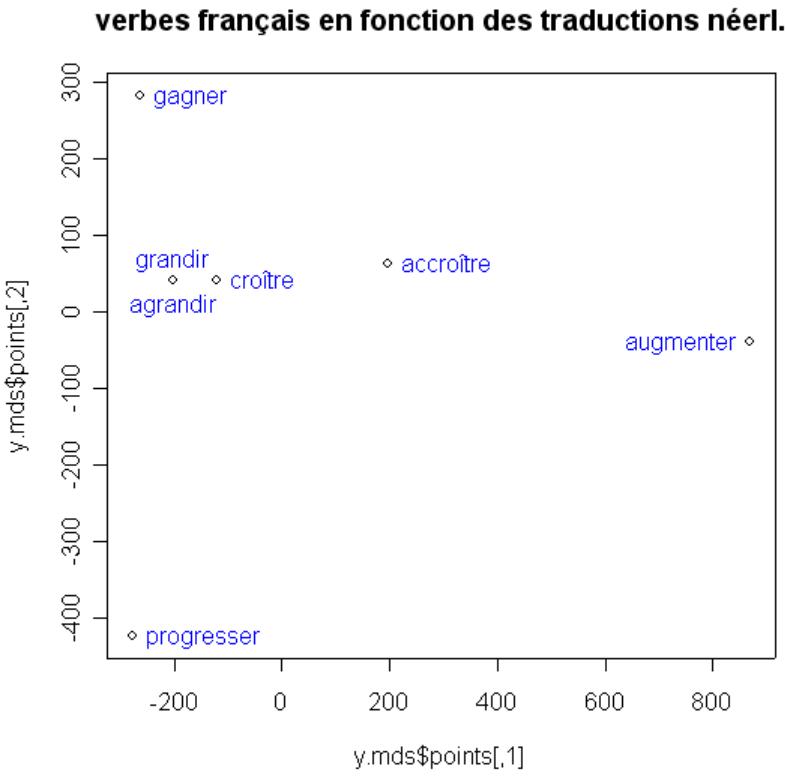


Figure 4 : Analyse MDS des verbes français en fonction de leurs traductions en néerlandais

Le profil de traduction particulier des trois pôles s'explique en partie par les fréquences élevées de certaines traductions privilégiées. Le retour aux fréquences des paires de mots nous apprend effectivement que le verbe *augmenter* se traduit de préférence par le verbe néerlandais *verhogen* (la fréquence de cette paire de mots s'élève à 775) et par le verbe *toenemen* (fréq. 550). *Progresser* se traduit principalement par *vooruitgang* (fréq. 320) et rarement par *toenemen* (fréq. 10). Le verbe *gagner* se traduit surtout par *winnen* (fréq. 464), ainsi que par *verdienen* (fréq. 133). Ces deux traductions privilégiées plus fréquentes témoignent du caractère polysémique du verbe *gagner*. Les dictionnaires de traduction devraient non seulement mettre en évidence les traductions privilégiées et fréquentes, mais ils devraient également fournir des indications de fréquence, par exemple sous forme d'astérisques (de 1 à 5).

La visualisation de l'analyse MDS des verbes français en fonction de leurs traductions en anglais confirme ces observations (Cf. figure 5). Les verbes *augmenter* et *progresser* se distancient des autres verbes en raison de quelques traductions privilégiées très fréquentes. Le verbe *augmenter* se traduit presque exclusivement par *increase* (fréq. 3293) et le verbe *progresser* se traduit surtout par *progress* (fréq. 640). Ces traductions ne sont pas attestées pour les autres verbes français ou moins souvent, ce qui permet d'expliquer la position périphérique des deux verbes commentés. En anglais, le verbe *gagner* n'a pas de traductions particulières très fréquentes ; il a un profil de traduction moins particulier et dès lors, il rejoint le nuage du groupe des autres verbes (Cf. figure 5).

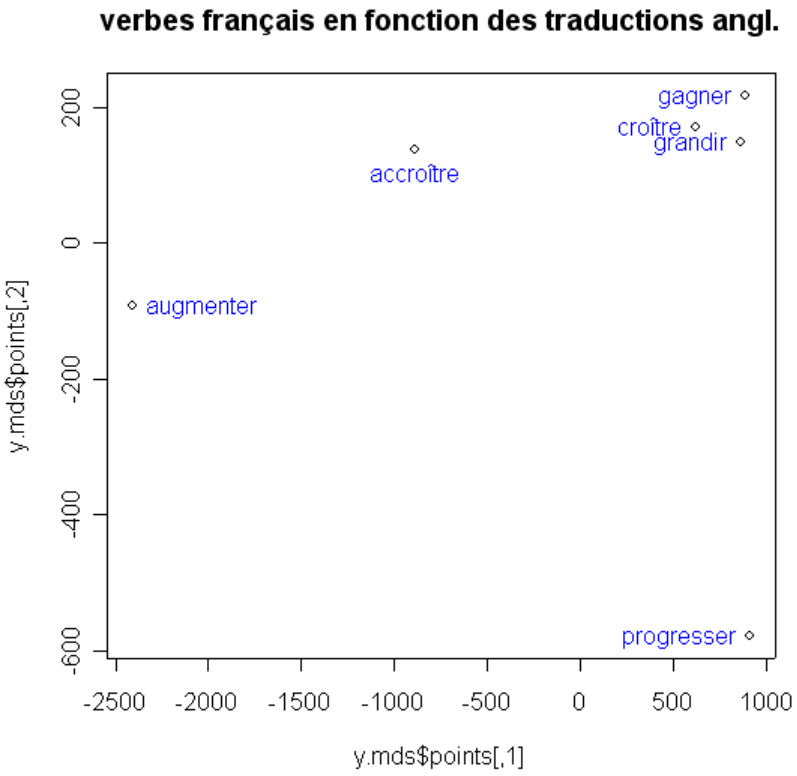


Figure 5 : Analyse MDS des verbes français en fonction de leurs traductions en anglais

4.2 Les profils combinatoires

Après les profils de traduction, nous avons également déterminé les profils combinatoires des verbes étudiés. Le logiciel WordSmith a permis d'identifier tous les collocatifs significatifs des verbes de hausse, en fonction de la mesure statistique log de vraisemblance (*log-likelihood ratio* ou LLR) et dans une fenêtre d'observation (*span*) de 5 mots à gauche et à droite. Il est à noter que les collocatifs dont la fréquence absolue est inférieure à 10 n'ont pas été inclus dans l'analyse des collocatifs.

A partir d'un sous-corpus spécifique par verbe de hausse, nous avons non seulement relevé les collocatifs significatifs, mais également les patrons (*patterns*) et les clusters de collocatifs prototypiques. Ces patrons et clusters permettent de réaliser une sorte de 'grammaire locale' lexicale (Gross, 1997) du verbe avec ses compléments (sujets, objets), prépositions et adverbes pertinents. La figure 6 visualise, à titre d'exemple, les emplois intransitifs du verbe *augmenter* dans la partie supérieure du tableau, ainsi que les emplois transitifs dans la partie inférieure. Le tableau montre que les collocatifs des verbes de hausse sont surtout des noms (sujets et objets). Ces tableaux de 'grammaire locale' sont particulièrement utiles pour les apprenants de FLE, qui ont souvent du mal à utiliser correctement les verbes et leurs compléments.

			prix	(ont)	légèrement	AUGMENTER	de			
prix	à	la	production		fortement		légèrement			
prix	à	la	consommation				fortement			
			exportations				rapidement			
			produits							
			salaires							
						AUGMENTER	sa	production		
							la			
							le	pouvoir	d'	achat
							le	prix		
								taux	d'	intérêt
								nombre	de	
								tarifs		
								offre		

Figure 6 : Collocatifs prototypiques du profil combinatoire du verbe augmenter

Les collocatifs pertinents, c'est-à-dire statistiquement significatifs, ont également fait l'objet d'une analyse MDS, dont le but était de caractériser les verbes de hausse les uns par rapport aux autres en fonction de leurs collocatifs. Est-ce que ces verbes ont des collocatifs en commun ? Est-ce que le profil combinatoire correspond au profil de traduction établi précédemment (Cf. section 4.1) ? L'analyse MDS des collocatifs des verbes de hausse a permis de retrouver les pôles *gagner* et *augmenter* et le nuage des autres verbes. Le verbe *gagner* se caractérise par un profil combinatoire très différent de celui des autres verbes, parce qu'il a des collocatifs pertinents très particuliers, notamment *argent*, *euros*, *manque* (dans *manque à gagner*), *terrain* (dans *gagner du terrain*). Les dictionnaires de traduction devraient mentionner cette variété de collocatifs et ces collocatifs privilégiés, mais surtout leurs contextes d'emploi, avec éventuellement des liens vers des phrases-exemples authentiques provenant de corpus monolingues disponibles en ligne.

Il ressort de ces analyses de corpus parallèles et monolingues que les profils de traduction des verbes français correspondent assez bien à leurs profils combinatoires. Les verbes *augmenter* et *gagner* se distinguent clairement des autres verbes, d'une part en raison de leurs traductions particulières et d'autre part en raison de leur collocatifs particuliers. Le verbe *augmenter* se caractérise par des traductions et collocatifs beaucoup plus fréquents que ceux des autres verbes, parce que c'est de loin le verbe le plus fréquent, tant dans les corpus parallèles que dans les corpus monolingues ciblés. C'est donc le verbe le plus important pour exprimer la notion de hausse en français. La position particulière du verbe *gagner* s'explique par sa polysémie, qui se reflète aussi bien dans quelques traductions fréquentes très différentes que dans les collocatifs sémantiquement hétérogènes.

4.3 Les profils combinatoires de quelques concepts-clés

Pour compléter l'analyse en termes de profils combinatoires, nous avons finalement étudié les moyens privilégiés pour exprimer la notion de hausse. A cet effet, nous avons repéré dans les corpus monolingues ciblés les collocatifs pertinents de quelques concepts-clés, c'est-à-dire de quelques phénomènes susceptibles à l'idée de fluctuation. Nous avons étudié les noms *bénéfice, cours, dollar, prix, taux, nombre, vente* pour le français, les noms *profit, dollar, price, percent, revenue, shares* pour l'anglais et les noms *winst, dollar, prijs, rente, aantal, aandeel, omzet* pour le néerlandais. Les collocatifs pertinents de ces concepts-clés expriment

la notion de fluctuation, c'est-à-dire la hausse ou la baisse. Les collocatifs ont été relevé dans WordSmith et soumis à une analyse MDS avec le logiciel R. Le but des analyses était de vérifier, d'une part, si la fluctuation s'exprime principalement par des verbes ou également par des mots appartenant à d'autres catégories grammaticales et s'il y a un patron préférentiel pour chaque langue et, d'autre part, s'il y a des collocatifs privilégiés par langue, c'est-à-dire des moyens privilégiés pour exprimer la notion de hausse.

Parmi les collocatifs des concepts-clés français, les substantifs *hausse* et *baisse* se distinguent clairement de tous les autres collocatifs qui expriment la notion de fluctuation (Cf. figure 7). Cette position particulière s'explique par leur co-fréquence très élevée avec les concepts-clés (co-fréquence de 762 pour *hausse* et de 474 pour *baisse*). En effet, leur co-fréquence est beaucoup plus élevée que la co-fréquence des verbes tels que *augmenter* avec les concepts-clés (co-fréquence de 50). Pour le français, les substantifs *hausse* et *baisse* sont donc les collocatifs privilégiés et le patron N + déterminant + N est de loin le plus fréquent et le plus important pour exprimer la notion de fluctuation (*la hausse des prix, la baisse du nombre*, etc.).

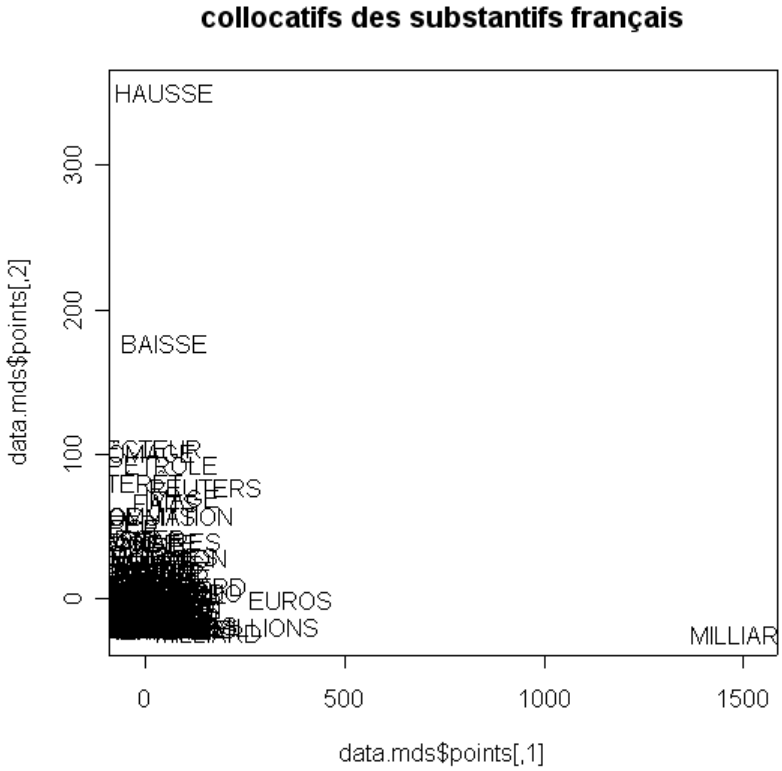


Figure 7 : Analyse MDS des collocatifs des 7 concepts-clés français

Par contre, les analyses de corpus montrent que les concepts-clés en néerlandais se combinent principalement avec des formes verbales (par exemple *stijgt, stijgen, steeg*) et déverbales (par exemple *stijgende prijzen*). Pour l'anglais, l'analyse des collocatifs des concepts-clés confirme la prépondérance des collocatifs verbaux (par exemple *rose*, *rises*) et déverbaux (par exemple *rising prices*). Elle indique également l'importance des adjectifs (par exemple *more*, *new*) et surtout des adverbes (par exemple *high, higher, up, down, more*), comme le montre la figure 8.

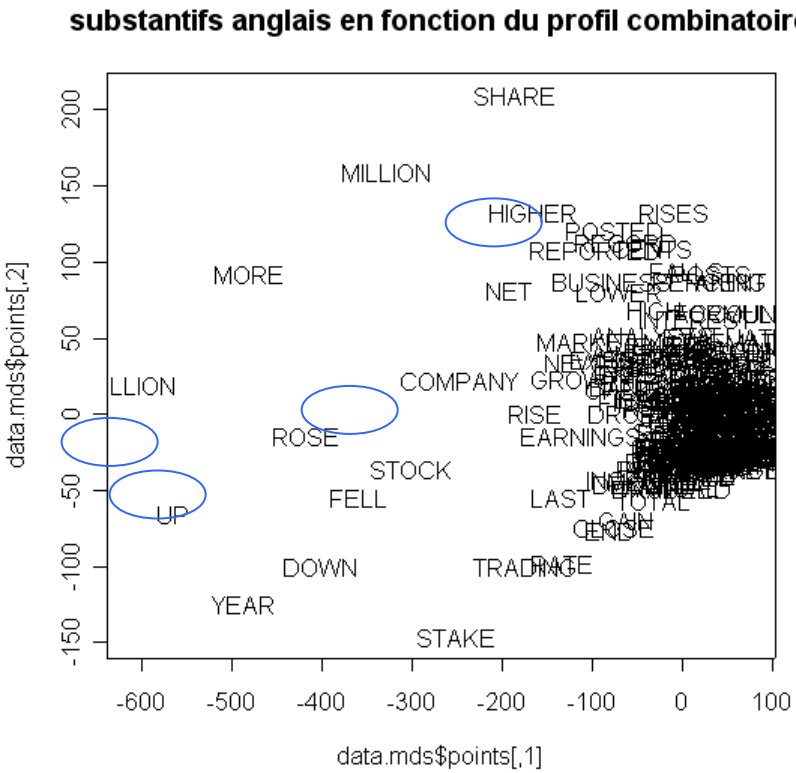


Figure 8 : Analyse MDS des collocatifs des 6 concepts-clés anglais

5 CONCLUSION

Cette étude exploratoire, dans le domaine de la lexicographie bilingue, a montré qu'il existe plusieurs manières d'enrichir les descriptions lexicographiques d'un dictionnaire de traduction à partir de nouvelles approches (analyses de fréquence et analyses MDS) basées sur l'analyse de corpus. Ces analyses aboutissent à une meilleure approximation des propriétés sémantiques et combinatoires des verbes étudiés, ce qui permet d'enrichir les descriptions lexicographiques traditionnelles des dictionnaires de traduction.

6 RÉFÉRENCES

Bertels A., Fairon C., Tiedemann J. et Verlinde S. (2009). « Corpus parallèles et corpus ciblés au secours du dictionnaire de traduction ». *Cahiers de Lexicologie*, vol. 94, p. 199-219.

Béjoint H. (2004). « Nouvelle lexicographie et nouvelles terminologies : convergences et divergences ». *Actes du Congrès de l'ACFAS.* Montréal.
http://archives.univ-lyon2.fr/154/1/bejoint_03.htm.

Fairon C. (2006). « Corporator: A tool for creating RSS-based specialized corpora ». *Proceedings of the Workshop Web as corpus. EACL 2006.* Trento.

Gross M. (1997). « The Construction of Local Grammars ». Dans E. Roche et Y. Schabes (éds.), *Finite State Language Processing,* Cambridge (Mass) : The MIT Press. p. 329–352.

Hundt M., Nesselhauf N. et Biewer C. (2007). *Corpus Linguistics and the Web*. Amsterdam/New York : Rodopi.

Kilgarriff A. et Grefenstette G. (2003). « Web as corpus : introduction to the special issue ». *Computational Linguistics*, vol. 29-3, p. 333-348.

Paumier S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique.* Thèse de Doctorat, Université de Marne-la-Vallée.

Rundell M. (1998). « Recent trends in English pedagogical lexicography ». *International Journal of Lexicography,* vol. 11-4., p. 315-342.

Tiedemann J. et Nygaard L. (2004). « The OPUS corpus - parallel & free ». *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbonne. http://stp.ling.uu.se/~joerg/paper/opus_lrec04.pdf.

Verlinde S. (1995). « La combinatoire du vocabulaire des fluctuations dans le discours économique ». *Cahiers de lexicologie*, vol. 66., p. 137-176.

ARGUMENTS LOGOMÉTRIQUES POUR LA DÉFINITION D'UNE DESCRIPTION EXTERNE DE CORPUS : APPLICATION AU CORPUS PIERRE MENDÈS FRANCE

Julien Bonneau
Laboratoire BCL – Université Nice Sophia-Antipolis
UMR 6039 du C.N.R.S

RÉSUMÉ

Cet article présente une méthode de classification de textes pour la définition de partitionnements de corpus. Basé sur des traitements statistiques endogènes des données textuelles et sur des critères de cohérence externe, nous appliquons notre approche à un corpus multiparamétré de discours de P. Mendès France et construisons des partitionnements pertinents pour nos critères, par genres textuels et par périodes chronologiques.

Mots clés : corpus multiparamétré ; descriptions externes ; genre textuels ; partitionnement endogène de corpus.

1 INTRODUCTION

L'accroissement des activités linguistiques sur corpus a ouvert un questionnement sur leurs descriptions externes, dynamisant, notamment, les travaux sur les typologies de textes, les situations d'énonciation et les genres de discours (Adam J.-M., 2004), (Rastier F., 2001). Autant d'éléments descriptifs souvent sollicités aussi comme paramètres de classification. Or cet usage ne manque pas de poser question à qui veut explorer un corpus textuel complexe et de grande ampleur : a) quelle valeur doit-on donner aux différentes typologies de textes proposées ? b) Comment décrire la situation d'énonciation et une description externe de celle-ci suffit-elle à rendre compte des contraintes langagières au niveau discursif ? c) Enfin, le genre de discours, qui relève d'une typologie par constat ouverte et mettant en jeu des descriptions dépassant largement les champs de la linguistique traditionnelle, peut-il être opératoire en linguistique ?

Cette contribution propose des éléments de réponse dans le but d'améliorer la description externe des corpus textuels multiparamétrés. Après avoir présenté le corpus et ses particularités, nous développerons notre méthode et ses principaux résultats sur la création de partitions en genres et diachronique du corpus.

2 LE CORPUS

Le corpus exploité dans cette étude est composé de 913 textes (13,5 Mo ; 1 500 000 occurrences) encodés selon les recommandations XML et TEI par le laboratoire ATILF de l'Université Nancy 2. Il correspond à la transcription électronique de (Mendès France P., 1984), sélection éditoriale de textes et discours de Pierre Mendès France entre 1922 et 1982, réalisée par François Stasse, Richard Dartigues et Simone Gros dans le but de rendre compte de « l'œuvre complète de Pierre Mendès France ». Par l'exploitation de travaux philologiques réalisés sur le corpus, nous avons enrichi celui-ci d'annotations spécifiques explicitant pour chaque texte original retenu pour (Mendès France P., 1984) : les caractères externes de ce texte tels que présentés dans (Biber D., 1988) ; son année de réalisation ; le/les genre(s) associé(s) par l'éditeur au texte.

Nos deux corpus de travail, formatés pour le logiciel Hyperbase (Brunet E., 2006), sont issus de ce corpus « père » et gardent son contenu textuel. Mais ils diffèrent dans leur

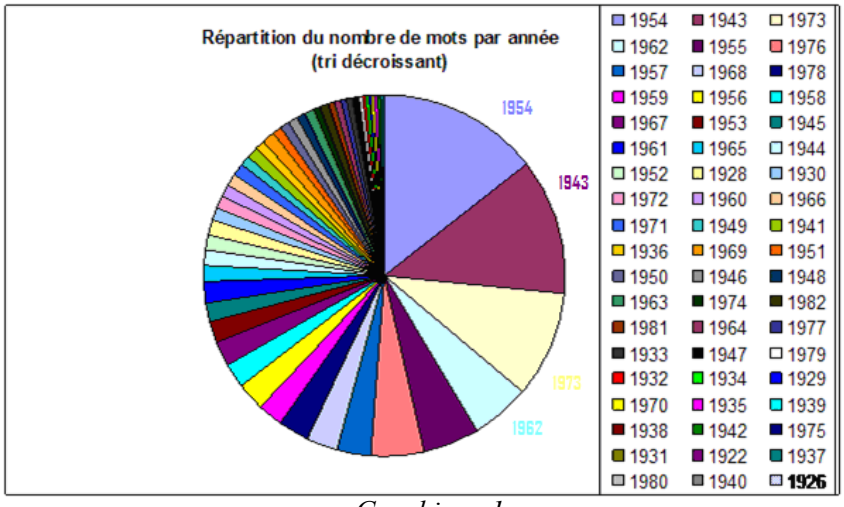
partition : l'une par années (57 parties) ; l'autre par genre éditorial et caractères externes de Biber (57 parties, 37 genres, croisés avec 16 sous-genres et les caractères de Biber).

2.1 Topologie du corpus

Pour un corpus multiparamétré d'une taille et d'une variété comme le nôtre, le profilage des parties est une étape de description de compréhension essentielle en termes de répartition quantitative et typologique. Nous avons réalisé, à l'aide d'Excel, différentes visualisations de la répartition des textes et des mots par genre (en différentiant les sous-genres) et par année. Quatre graphiques en camembert (*graphiques 1*, *2*, *3* et *4*) représentent respectivement la répartition quantitative des mots et des textes par genre et par année. Un dernier histogramme (*graphique 5*) permet la visualisation relative, en pourcentage, de la répartition des genres par année.

Nous utilisons ces graphiques comme accès à deux types d'informations : a) d'une, part un point de vue diachronique de la répartition quantitative des mots et des textes, et par inférence du discours, de Pierre Mendès France ; b) d'autre part, la répartition et l'évolution quantitative diachronique des textes entre les genres de discours de Mendès France. Pour compléter et affiner ces observations, nous effectuons des retours aux données bruts, présentées sous forme de tableaux croisant années et genres (non présentés ici car trop grand, 57x57), notamment pour le repérage des genres du corpus textuellement sous-représentés et donc non-satisfaisants en tant que définition de sous-corpus. Nous portons une attention particulière à ces genres, car ils définissent des parties du corpus candidates privilégiées aux regroupements avec d'autres genres, sous peine d'être non-utilisables et donc écartées de l'analyse.

Nous présentons dans la suite les principaux résultats de cette première approche du corpus :

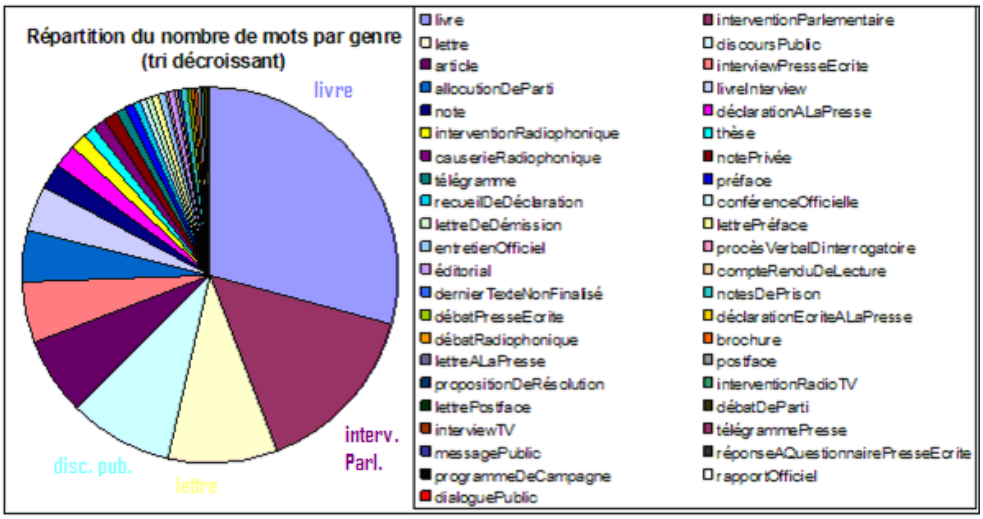


Graphique 1

Ce premier camembert présente la répartition du nombre de mots du corpus par ordre décroissant en fonction des années. Cette répartition est très hétérogène, comme le montre le simple constat qu'à six années, soit à peu près 1% des années présentent dans le corpus, 1954 (15%), 1943 (12%), 1973 (9%), 1962 (5%), 1955 (5%), 1973 (5%), correspondent plus de 50% des mots du corpus et qu'aux 41 dernières années de ce classement, soit un peu moins de 72% des années du corpus, correspondent seulement 26% des mots du corpus.

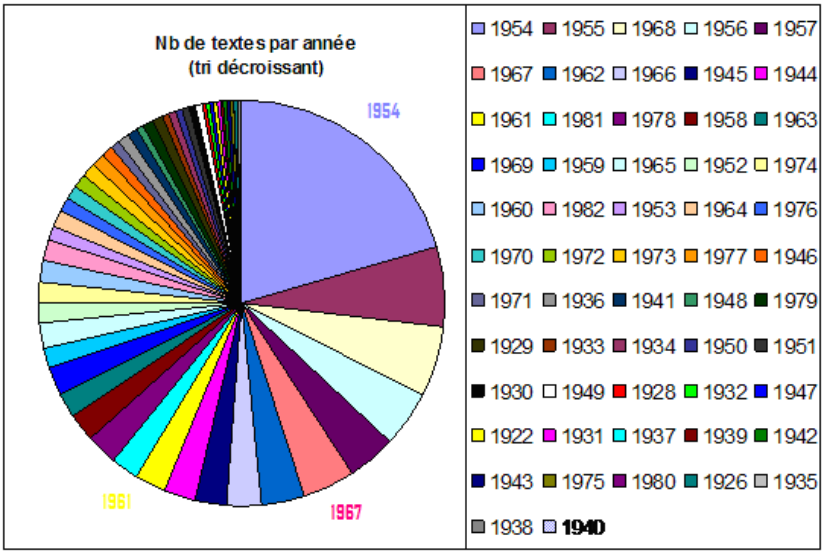
Le même constat s'impose pour les genres, comme le montre le *graphique 2* (genres et sous-genres confondus, chaque sous-genre ayant été distingué de son genre dans le

comptage). Les 4 principaux genres en nombre de mots du corpus (moins de 1% des genres) monopolisent 62% des mots du corpus, le genre *livre* représentant à lui seul près de 30% des mots. Les 30 derniers genres triés par ordre décroissants de mots, soit plus de 50% des genres du corpus, correspondant à moins de 10% des mots.



Graphique 2

La différence de répartition des mots entre les parties du corpus, caractérisées par année ou par genre, amène à deux précautions statistiques afin de permettre leur comparaison : a) privilégier le travail sur les codes grammaticaux et les structures, pour limiter la diversité linguistique (notamment thématique) statistiquement plus forte sur un grand nombre de formes, restreignant ainsi le nombre d'unités linguistiques observables (les lemmes seront néanmoins utilisés pour préciser la construction de la partition diachronique, nécessairement soumise aux évolutions thématiques) ; b) préférer, dans nos analyses, les fréquences relatives ou la présence/absence d'une unité linguistique aux fréquences absolues pour limiter l'effet de « poids statistique » du nombre d'occurrences dans les différentes parties du corpus.

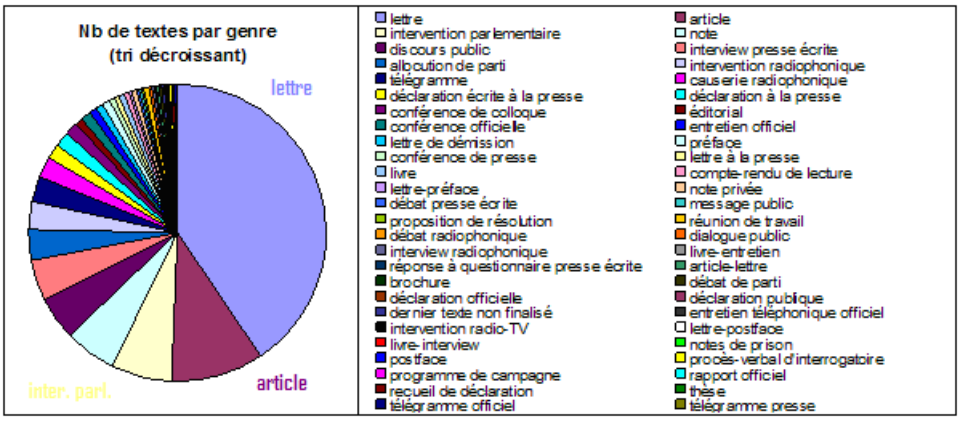


Graphique 3

Le *graphique 3* montre l'hétérogénéité de la répartition des textes à travers les années : a) l'année 1954 représente 21% des textes du corpus, aucune autre année ne dépassant les 6% ;

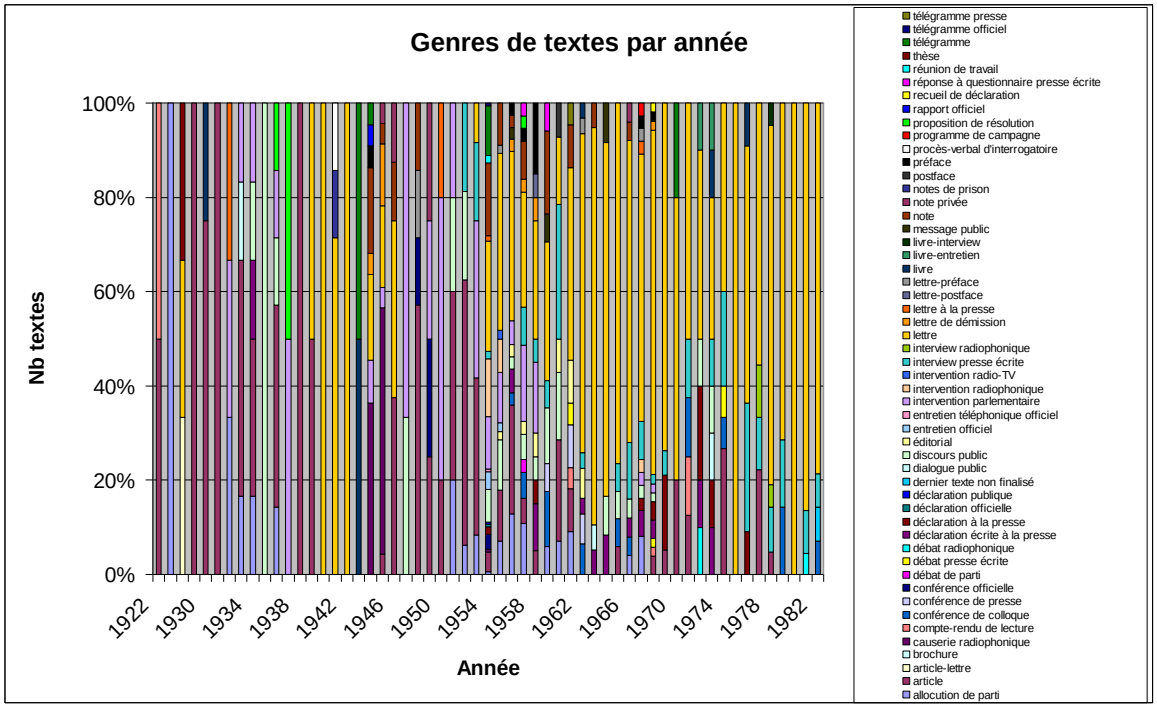
b) les 17 années les moins fournies en nombre de textes comprenant chacune moins de 1% des textes du corpus, pour un total de 5% des textes. Dernières de ce classement, les années 1926, 1935,1938 et 1940 ne correspondent qu'un seul texte.

Il en va de même pour la répartition des textes à travers les genres (*graphique 4*), où les écarts sont encore plus marqués : a) le genre *lettre* contient 40% des textes du corpus ; b) article 10% ; les 29 genres contenant les moins de textes du corpus, soit chacun moins de 1% des textes, pour un total cumulé de 4% des textes du corpus. Parmi ces 29 genres, les 19 derniers ne sont représentés que par un texte.



Graphique 4

Ce constat met en valeur la nécessité de regroupements des parties du corpus entre elles, pour s'assurer de leur représentativité au niveau textuel. Les candidats regroupements privilégiés seront donc les genres et les années contenant un très faible nombre de textes, sous peine d'être exclues de l'analyse.



Graphique 5

Ce dernier graphique (*graphique 5*) présente la répartition chronologique des genres à travers le corpus. Il met à jour un glissement dans l'activité discursive de Pierre Mendès

France. La première moitié du graphique, de 1922 à 1961, qui correspond à la période parlementaire de l'activité politique de Mendès France, présente une forte proportion de textes de genres article, allocution de parti et intervention parlementaire, sauf de 1940 à 1942 (Seconde Guerre Mondiale). La deuxième moitié, de 1962 à 1982, est dominée par le genre lettre et voit aussi apparaître une forte proportion du genre interview presse écrite.

Cet histogramme permet de décrire une forme d'évolution du discours de Mendès France à travers le temps et donne, par là, de premiers indices de partitions diachroniques possibles du corpus.

2.2 La méthode

Pour construire une typologie aussi endogène que possible du corpus, nous confrontons la typologie diachronique, la typologie de (Biber, 1988) et la typologie des genres éditoriaux à un traitement statistique des données textuelles. Nous utilisons pour nos analyses la version lemmatisée d'Hyperbase 8.0, qui intègre le logiciel Tree Tagger de l'Institute for Computational Linguistics de l'Université de Stuttgart et son jeu d'étiquettes, nous permettant donc de travailler sur différentes unités linguistiques : les formes, les lemmes, les catégories grammaticales et les structures grammaticales. Notre méthode s'appuie sur la fonction Distribution du logiciel Hyperbase et son utilisation, pour la visualisation, des analyses arborées, théorisées par X. Luong (Luong X., 1994). La fonction Distribution fournit la matrice de la distance intertextuelle (sur les fréquences relatives) entre les différentes parties du corpus. A partir de celle-ci, l'analyse arborée propose une représentation graphique de l'intégralité de son contenu informationnel (*graphiques 6* à *9*).

L'articulation contrastive des résultats d'analyses des différents corpus, ainsi que l'observation de la répartition diachronique des genres, nous permettent de rationaliser, sur des arguments statistiques, les meilleurs regroupements de textes possibles à partir de nos étiquettes de départ. Le but est d'obtenir une profondeur de description optimale pour le corpus et d'améliorer ainsi son partitionnement et sa description externe. Le traitement est ainsi réitéré, à l'aide de la fonction Genre d'Hyperbase, qui permet le regroupement et le reétiquetage de plusieurs parties du corpus sous un même genre, jusqu'à ce que les arguments externes de regroupements ou de dégroupements soient épuisés. Ainsi, on obtient pas à pas des regroupements endogènes qui affinent la description externe du corpus, tout en limitant le nombre d'étiquettes utilisées.

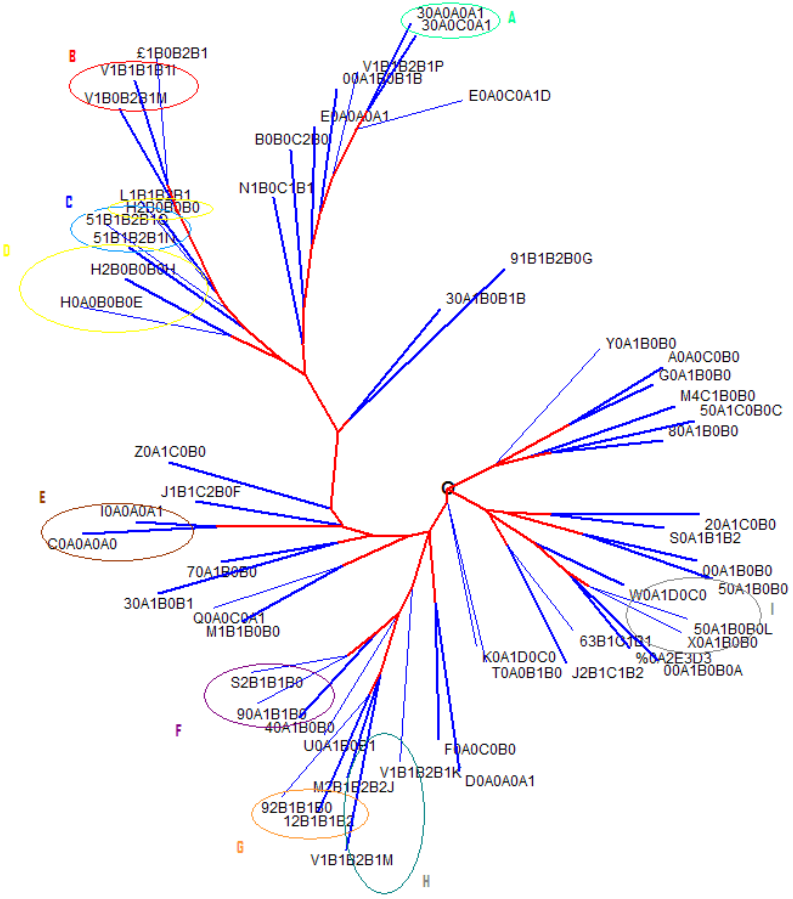
Pour les genres de discours les arguments de regroupements et de dégroupements peuvent être de 3 sortes : a) retenir un sous-genre ou son genre ; b) (re-)spécifier un genre si un caractère de Biber apparaît comme discriminant pour celui-ci à l'intérieur d'un genre initial. Les caractérisations de Biber ne sont donc utilisées que pour spécifier certains types de regroupements des parties du corpus ; c) regrouper plusieurs genres éditoriaux jugés connexes par l'analyste sous un seul genre.

3 RÉSULTATS : Les genres à travers les textes

3.1 Premières résultats

Pour affiner l'observation des genres, nous avons réalisé un découpage du corpus qui croise genres, caractères externes de Biber et sous-genres. Ces informations sont explicitées dans l'étiquetage des parties du corpus, balisées à l'aide de 8 à 9 caractères (suivant l'appartenance ou non à un sous-genre), chacun correspondant, dans l'ordre : a) à un genre ; b) aux caractères de Biber ; c) éventuellement à un sous-genre. Cette annotation doit nous permettre de nous repérer dans la lecture des analyses arborées, afin de faciliter les regroupements en genre. On obtient, *in fine*, une partition des données en 57 sous-corpus point de départ de notre analyse.

Dans cet article, nous limitons nos analyses arborées au calcul de distance sur la fréquence relative des différents codes grammaticaux en fonction des parties du corpus.

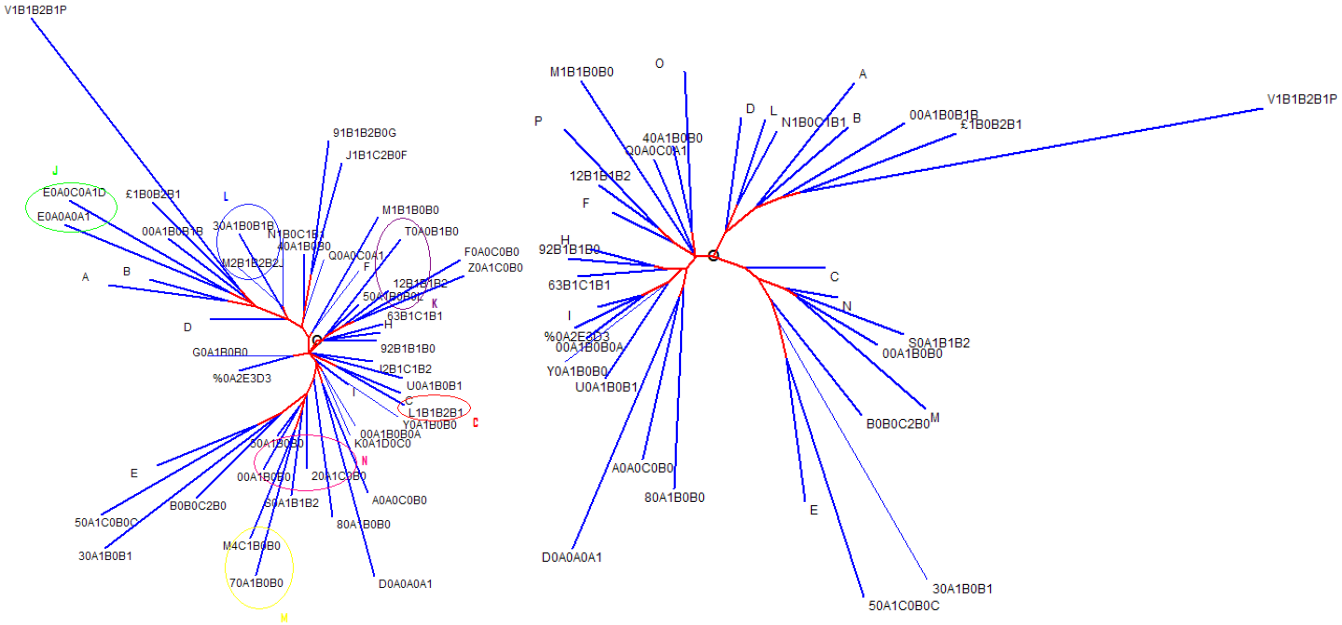


Graphique 6 : Analyse arborée des codes grammaticaux du corpus genre/Biber

Le *graphique 6* présente les résultats de notre première analyse arborée des genres textuels. Entouré en rouge, nous avons regroupés sous l'étiquète **B**, à l'aide de la fonction Genre d'Hyperbase, les débats de parti et les débats radiophoniques, qui possèdent une bonne proximité dans l'arbre. Nous avons réalisé la même opération pour les lettres décrites selon les caractères de Biber comme privées et institutionnelles (**A**, cerclés en vert), les livres d'entretiens et les livres d'interviews (**C**, en bleu), les différents sous-genres d'interventions radiophoniques (**D**, en jaune), les notes de prison et les notes privées (**E**, en marron), etc.

Procédant de manière itérative, nous avons réalisé une nouvelle analyse arborée du corpus ainsi modifié, dont le résultat est l'arbre de gauche du *graphique 7*, où de nouveaux regroupements, toujours entourés dans le graphique, se sont proposés à nous. Nous avons répété l'opération 6 fois en tout avant d'épuiser les regroupements proposés par cette méthode, l'arbre final étant présenté dans la partie droite du *graphique 7*.

A ce stade nous avons réduit les parties du corpus de 57 à 35, ce qui correspond à peu près au nombre de genres de départ (37) de nos analyses et à plus de 50 % de regroupements des genres ou sous-genres représentant 1% ou moins du corpus (*cf. supra*, *graphique 4*). Nous avons donc réorganisé genres et sous-genres sur des rapprochements de caractères externes et de critères statistiques. Les regroupements obtenus sont présentés dans le *tableau 1*. Les regroupements issus de la première analyse sont colorés en rouge dans le tableau, de la deuxième analyse en bleu, de la troisième en vert, de la quatrième en mauve, de la cinquième en marron (des crochets sont utilisés pour matérialiser les regroupements effectués entre regroupements déjà existants).



Graphique 7 : Regroupements par l'analyse arborée itérative des codes grammaticaux du corpus genre/Biber

Ayant épuisés les regroupements statistiques, nous nous intéressons ensuite à la cohérence externe du *tableau 1* et de la dernière représentation arborée. Deux options théoriques peuvent-être envisagées : regrouper des genres très éloignés dans l'arbre mais de caractérisation cohérente, ou exclure les genres trop peu représentés. Dans la suite, nous regroupons certains des regroupements de genre effectués avec des genres restant pour construire de nouveaux groupes plus hétérogènes sur nos critères statistico-linguisitiques, mais satisfaisants du point de vue de la description situationnelle des textes.

Nous commençons par l'observation du *tableau 1* :

- Le groupe **A** correspond aux **lettres et notes de destinataire unique**, privé ou institutionnel.
- Les groupes B et H correspondants au genre **Débat** sont liés par les Débats radiophoniques dans le seul caractère externe discriminant tient en leur publication, l'un l'étant, l'autre non. Cette discrimination nous semblant trop faible, nous avons décidé de les regroupés dans le groupe **B**.
- Les **interviews et entretiens** sont regroupés dans le groupe **C**.
- Au groupe **D** sont associés les **discours radiophoniques**.
- Les **notes personnelles de Pierre Mendès France** forment le groupe **E**.
- Le regroupement M nous fait inférer que le sous-genre Déclaration à la presse statut inconnu appartient en réalité au sous-genre Déclaration écrite à la presse. De l'observation des regroupements L, M, F et O nous retirons une nouvelle tripartition, d'une part le regroupement L et M, groupe **L** qui correspond à des **textes écrits destinés à la presse pour une diffusion publique** (sur support écrit ou oral), d'autre part le regroupement **F** qui correspond aux **discours à support écrit directement destinés à un public**, enfin le regroupement **O** associé aux **déclarations orales non écrites**.
- Au groupe **I** correspond les **préfaces et postfaces**.
- Les **ouvrages divers non officiels** sont regroupés en **N**.

- A **P** correspondent les **ouvrages divers officiels**.

Sur ces critères, on obtient l'analyse arborée du *graphique 8* (partie de gauche).

A	B	C	D	E	F	H
[[-Lettre privées -Lettres institutionnelles] [-Lettre de démission -Note avec interaction]]	-Débat de parti -Débat radiophonique non-publié	-Livre entretien -Livre interview -Interview presse écrite	-Intervention radio-TV -Causerie radiophonique -Intervention radiophonique	-Note privés -Notes de prison	[[-Discours public écrit -Conférence de colloque écrit-lu] [-Conférence officielle écrit-lu -Message public écrit -Recueil de déclarations écrit]]	-Débat presse écrite -Débat radiophonique publié

I	L	M	N	O	P
-Postface -Lettre postface -Préface -Lettre préface	-Conférence de presse écrit-lu -Lettre à la presse	-Déclaration à la presse statut inconnu -Déclaration écrite à la presse	-Livre -Thèse	-Déclaration officielle oral -Déclaration public oral	-Rapport officiel -Programme de campagne

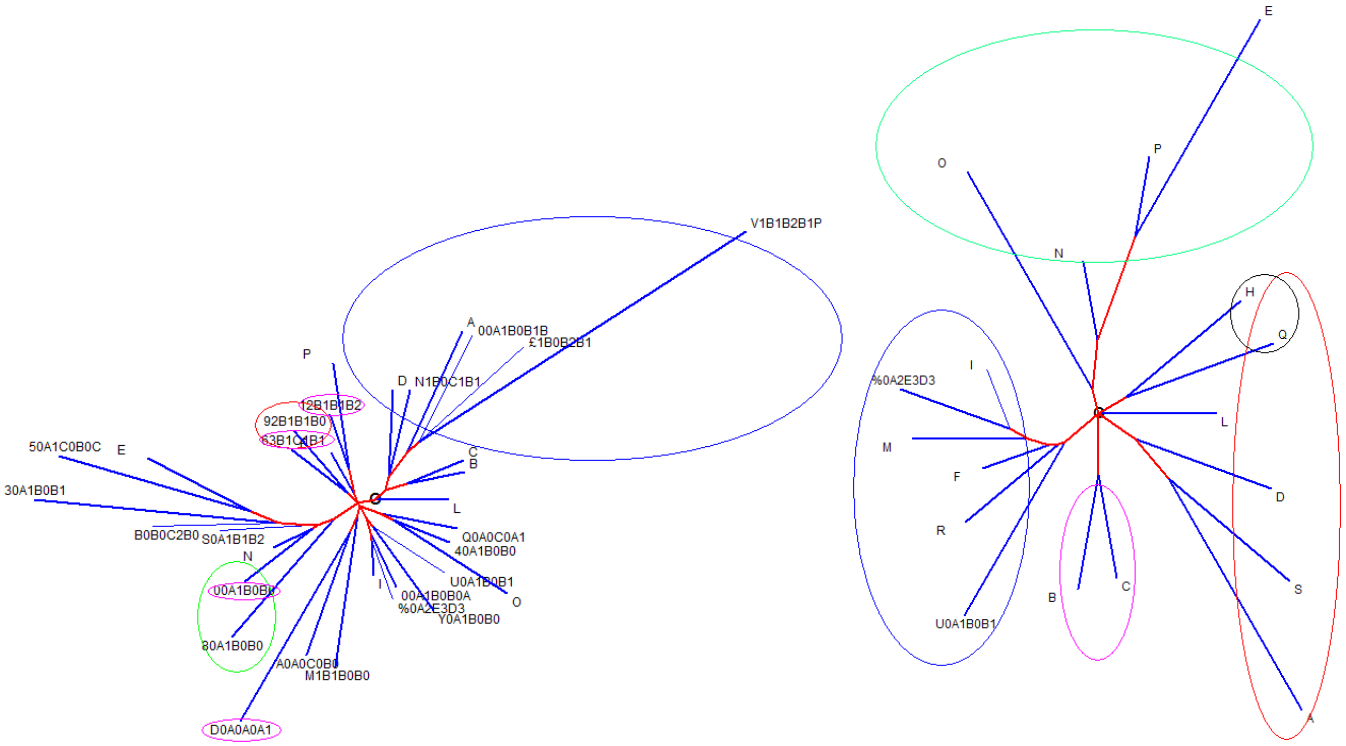
Tableau 1 : Regroupements construits par analyses arborées

A partir de ce nouvel arbre, nous posons à nouveau la question de la cohérence dans la nouvelle partition de corpus obtenue, en tenant compte des positions des genres et groupes de genre dans l'arbre. Certains regroupements cohérents sur critères statistiques et externes sont apparus. C'est le cas pour : le groupe F et les discours public écrit-lu (F + 9…, entourés en rouge) ; les articles et les éditoriaux (0…0 + 8…, entourés en vert), regroupement **M**.

Les genres décrient comme représentatifs sur critères statistiques ne nécessitent pas de regroupements (*cf. supra*, *graphiques 2* et *4*, entourés en mauve dans l'arbre). C'est le cas des : allocutions de parti (1…, groupe **H**) ; interventions parlementaires (6…, groupe **Q**) ; articles (0…0) ; des télégrammes (D…1).

La branche supérieure droite de notre arbre regroupe l'ensemble des textes avec forte interaction au sens de Biber (entourée en bleu) : entretiens, débats, interviews, dialogues. Nous décidons de prendre les genres correspondant isolés dans cette position (dialogue public V…P, interview radiophonique £…, entretien officiel N…) pour construire un regroupement **S,** textes divers avec forte interaction.

Pour des raisons de cohérence externe des regroupements nous complétons certains d'entre eux. C'est le cas pour : le genre conférence de colloque écrit et le regroupement **F** ; le regroupement **L** et les sous-genres lettre-article, lettre publique et lettre à la presse ; télégramme et ses sous-genres, **R** ; le sous-genre déclaration orale à la presse et le regroupement déclarations orales, **O** ; les genres et sous-genres brochure (officielle), procès-verbal d'interrogatoire, proposition de résolution et le regroupement **P**, ouvrages divers officiels ; le sous-genre compte-rendu de lecture et le regroupement **M**, article.



Graphique 8 : Analyses arborées après groupement sur critère externe des regroupements itératifs et analyse arborée des regroupements finaux

Reste à le fin de cette analyse les genres *dernier texte non-finalisé* et *réponse à un questionnaire presse écrite*, tous deux sous-représentés (2 textes) et donc écartés des analyses.

Nous reportons les regroupements effectués dans le *tableau 2*, les dernières modifications étant reportées en rouge.

Tableau 2a : Les regroupements en genre finaux

A	B	C	D	E	F	H
-Lettre privées -Lettres institutionnelles -Lettre de démission -Note avec interaction	-Débat de parti -Débat radiophonique non-publié -Débat presse écrite -Débat radiophonique publié	-Livre entretien -Livre interview -Interview presse écrite	-Intervention radio-TV -Causerie radiophonique -Intervention radiophonique	-Note privés -Notes de prison	-Discours public écrit -Conférence de colloque écrit-lu -Conférence officielle écrit-lu -Message public écrit -Recueil de déclarations écrit -Discours publique écrit-lu -Conférence de colloque écrit	-allocution de parti
I	**L**	**M**	**N**	**O**	**P**	**Q**
-Postface -Lettre postface -Préface	-Conférence de presse écrit-lu -Lettre à la	-Article -Editorial -Compte-rendu de	-Livre -Thèse	-Déclaration officielle oral -Déclaration public oral	-Rapport officiel -Programme de campagne	-Intervention parlementaire

-Lettre préface	presse non-publiée -Déclaration à la presse statut inconnu -Déclaration écrite à la presse -Lettre-article -Lettre publique	lecture		-Déclaration orale à la presse	-Brochure -Procès-verbal d'interrogatoire -Proposition de résolution	
R	**S**					
-Télégramme -Télégramme à la presse -Télégramme officiel	-Dialogue public -Interview radiophonique -Entretien officiel					

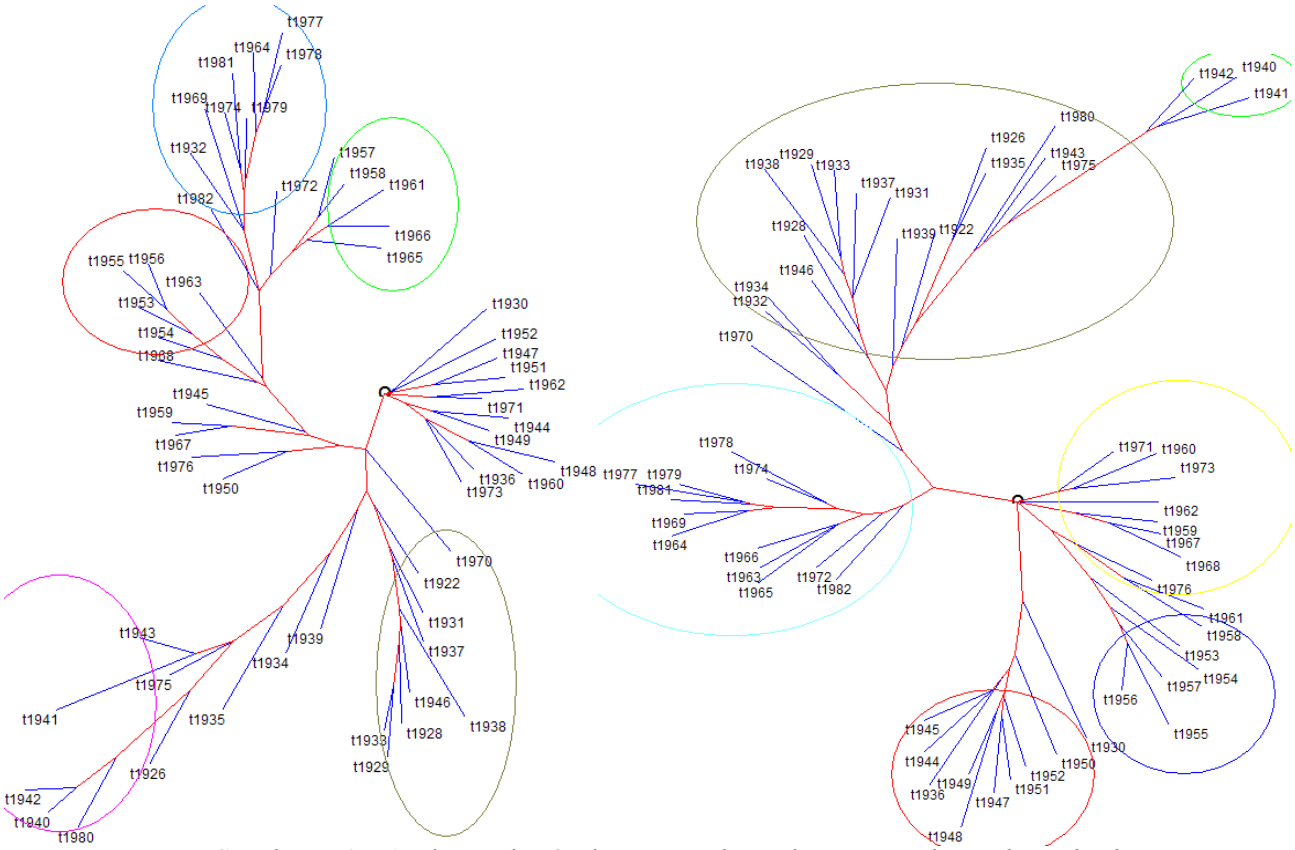
Tableau 2b : Les regroupements en genre finaux

L'observation de ce tableau met en valeur l'hétérogénéité des regroupements paradigmatiques effectués : oral *vs*. écrit ; différentiations médiatiques ; différentiations situationnelles et interactionnelles ; etc. L'observation par analyse arborée de ces regroupements (*graphique 8* partie de droite) montre néanmoins que cette hétérogénéité s'organise en des pôles bien discriminés autours de ces différentes axialités : écrit (bleu et vert) ; oral (rouge) ; forte interaction (mauve) ; allocution politicienne (noire).

3.2 Second resultats : Regroupements diachroniques des textes

Nous appliquons la même méthode sur la partition par années de notre corpus. Dans cette partie nous appuierons nos regroupements sur les codes, mais aussi sur les lemmes, afin de tenir compte de l'influence thématique, liée, en partie, au contexte historique et donc à la chronologie. Nous présentons dans le *graphique 9* les résultats de l'analyse arborée de ce corpus sur les codes (à gauche) et sur les lemmes (à droite).

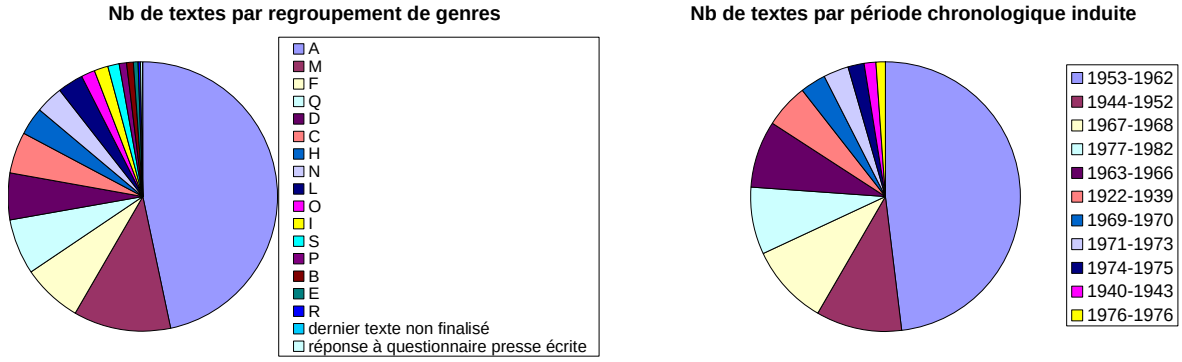
Les regroupements apparaissent dès ce niveau de manière assez nette, particulièrement au niveau lemmatique : Les premiers textes (1922-1939) ; la seconde guerre mondiale (1940-1943) ; l'après-guerre et sa préparation (1944-1952) ; la période parlementaire (1953-1962) ; des années soixante à quatre-vingt, avec des dates particulières correspondant à des retours ponctuels à la vie politique (en jaune) : 1967-1968, 1971, 1973, 1976.



Graphique 9 : Analyse arborée du corpus chronologique sur les codes et les lemmes

4 CONCLUSION : Topologie induite du corpus

Les traitements de regroupements effectués sur le corpus ont modifié en profondeur sa structure interne comme sa description externe. En contre point de notre première « Topologie du corpus » (*cf. supra*), nous proposons, en conclusion, la description de la nouvelle topologie induite du corpus.

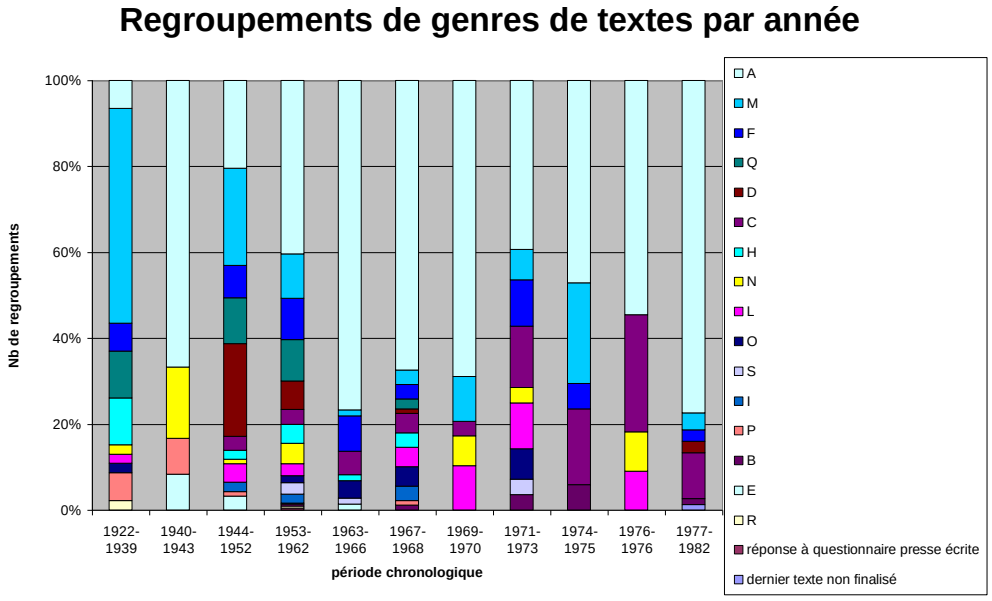


Graphique 10

Le *graphique 10* présente les profils de la répartition des textes en genre (à gauche) et par période (à droite). Dans les deux partitions on voit apparaître le même type de phénomènes : a) la création d'un regroupement presque majoritaire ; b) la diminution des écarts en nombre de textes entre les autres regroupements. Satisfait de la diminution de ces déséquilibres (b), nous envisageons de traiter les phénomènes (a) par deux approches différentes. Pour les regroupements de genres, nous exclurions les lettres privés du regroupement A, car c'est ce genre, que nous jugeons peu significatif au niveau discursif pour un personnage publique

comme P. Mendès France, qui apporte le déséquilibre (*cf. graphique 4*, *supra*). Pour le découpage chronologique, Nous redécouperions la période 1953-1962 par années ou paires d'années pour palier au déséquilibre construit.

Nous présentons enfin, dans le *graphique 11*, la nouvelle répartition des genres à travers le découpage chronologique, où l'on retrouve les tendances décrites pour le *graphique 5* (*cf. supra*).



Graphique 11

Le principal avantage de notre approche est de ne pas imposer au corpus un découpage sociolinguistique strict *a priori*. Elle donne, de plus, un appareil critique sur la construction des partitionnements obtenus, notamment en termes de types d'arguments (statistiques / critères externes) et de profondeur statistique des regroupements. C'est sur des arguments d'abords statistiques, puis externes que nous obtenons de nouvelles partitions cohérentes du corpus, chacun de ces arguments étant à la fois explicite et critiquable.

5 RÉFÉRENCES

Adam J.-M. (2004). *Linguistique textuelle : des genres de discours aux textes*. Paris : Nathan.

Biber D. (1988). *Variaton across speech and writing*. Cambridge : Cambridge University Press.

Brunet E. (2006). *Hyperbase, Manuel de référence*. Nice : UNSA.

Luong X. (1994). *L'analyse arborée des données textuelles : mode d'emploi*. Nice : UNSA.

Luong X. et al. (2003). *La distance intertextuelle* in *Corpus, n°2*. Nice : UNSA.

Mendès France P. (1984). *Pierre Mendès-France. Oeuvres Complètes*. Paris : Gallimard.

Rastier F. (2001). *Arts et sciences du texte*. Paris : PUF.

ONTOLOGIES NATURELLES ET COERCION : FORMALISATION DE CONNAISSANCES À PARTIR D'OBSERVATIONS EN CORPUS

Ismaïl El Maarouf[1], Marc Le Tallec[2] et Jeanne Villaneau[3]
[1] Laboratoires HCTI-Licorn, Valoria – UBS-UEB
[2] Laboratoire LI – Université de Tours
[3] Laboratoire Valoria – UBS-UEB

RÉSUMÉ

Le projet EmotiRob a pour finalité de créer un robot compagnon interactif destiné à des jeunes enfants fragilisés de 3 à 7 ans, comme par exemple dans le cas d'une longue hospitalisation. Le robot est une peluche conçue pour être tenu dans les bras, potentiellement capable d'émettre quelques sons simples et d'exprimer des émotions primaires. Le but est qu'il réagisse aux propos de l'enfant de façon aussi appropriée que possible.

L'implémentation d'un tel robot requiert un module de « compréhension » des propos de l'enfant dont le but est de détecter les concepts présents dans l'énoncé, ainsi que les relations sémantiques qui les unissent. Pour réaliser la « compréhension » des énoncés, nous avons choisi d'adapter le système LOGUS, mise en œuvre d'une approche logique permettant une compréhension robuste sur des tâches complexes. LOGUS n'utilise pas de patrons sémantiques prédéfinis : les associations possibles entre concepts reposent sur une connaissance sémantique qui autorise ou non leur regroupement.

Notre objectif est de construire une ontologie naturelle à destination du module de compréhension. Elle repose sur la définition d'une hiérarchie de types sémantiques. Chaque concept est classé dans un ou plusieurs types en fonction des relations sémantiques que l'étude des collocations permet d'observer en corpus.

Nous avons constitué un corpus de contes qui contienne une variété de concepts et de relations sémantiques propres à l'univers enfantin. Le corpus EmotiRob comporte 138 contes pour un total de 160 000 mots. Il est de taille relativement réduite mais constitue un point de départ satisfaisant pour l'objectif qui nous importe. Toutes les relations syntaxiques entre verbe et nom ont été extraites pour les verbes ayant une fréquence supérieure à 30 (130 verbes).

L'analyse des listes d'unités lexicales (ayant la même fonction syntaxique vis-à-vis d'un verbe) montre d'une part que le sens du verbe co-varie avec le type sémantique de ses arguments, et d'autre part, avec quels types sémantiques un verbe peut se composer dans le cadre d'une relation syntaxique donnée, mettant ainsi en lumière les phénomènes de coercion, dans lesquels le type attendu par un prédicat entre en conflit avec le type observé.

Nous avons cherché à modéliser et rendre compte de phénomènes comme la coercion pour des associations sémantiques attestées telles que « boire un verre », « entendre la voiture » ou « passer les dragons » (considérés en l'occurrence comme des obstacles). Nous avons dû adjoindre à notre hiérarchie de types deux outils supplémentaires : la conversion de types et les propriétés saillantes. Par exemple, la conversion du type véhicule en type son (tout véhicule produit effectivement du bruit) permet à une voiture d'être entendue; préciser que l'une des propriétés saillantes d'un verre est d'être utilisé pour boire (dit rôle « TELIC ») permet de rendre possible l'analyse de « boire un verre ». Joints à la hiérarchie de types sémantiques, ces connaissances permettent une couverture satisfaisante des relations sémantiques effectivement observées, sans surgénération effective.

1 INTRODUCTION

Cette étude se situe aux confluents de la linguistique et de l'informatique, terrain déjà occupé par la linguistique de corpus et le traitement automatique des langues. Ces disciplines oeuvrent pour l'échange et la collaboration de données, modèles et théories qui puissent converger vers des descriptions de dimensions encore inexplorées de la langue. Cet article se veut être une modeste contribution aux recherches sur le plan sémantique du texte et a pour objectif de mettre en lumière les modèles sémantiques par rapport aux données attestées collectées en corpus. Le problème a émergé récemment et se traduit par des prises de position différentes: d'une part ceux qui comme J. Pustejovsky (Pustejovsky & Ježek, 2008), cherchent à valider leur modèle théorique à travers des analyses de corpus et d'autre part, ceux qui, comme P. Hanks (Hanks, 2008), cherchent à établir des modèles à partir du corpus. Bien qu'elles soient présentées comme opposées, ces deux approches partagent en fait de nombreuses affinités et c'est vis-à-vis du traitement d'un phénomène particulier qu'elles diffèrent.

La recherche ici présentée est une analyse de corpus qui reprend le modèle de Hanks et s'intéresse au phénomène de coercion et d'alternance de type sémantique. Le problème qui se pose est de savoir comment gérer des occurrences langagières, communément désignées comme métonymiques, telles que « boire un verre » par rapport à « boire de l'eau » ou encore « se rendre à la messe » par rapport à « se rendre à l'église » car « verre » et « eau » sont de nature ontologique différente (tout comme le sont « messe » et « église ») et ce phénomène doit pouvoir être résolu par le modèle sémantique ou système informatique lorsqu'il y a lieu.

Cette analyse a émergé d'un contexte applicatif que nous décrirons dans un premier temps. Elle s'insère d'une part dans le cadre d'un projet de recherche interdisciplinaire, plus particulièrement le module de compréhension d'un robot-compagnon, et d'autre part vis-à-vis d'un système de compréhension de l'oral spontané qui repose sur une « connaissance sémantique » du domaine visé.

L'étude s'appuie sur cette « connaissance » en la confrontant au corpus, c'est pourquoi nous présenterons dans un second temps les données de l'analyse, c'est-à-dire le corpus, le modèle sémantique employé ainsi qu'une description plus fine du phénomène d'alternance de type sémantique en fonction des occurrences rencontrées en corpus.

Enfin, nous ferons état des résultats de l'expérience menée sur le système sémantique, les types d'alternances rencontrées, ainsi que les solutions proposées pour intégrer les observations de l'analyse de corpus au système, sans pour autant boulverser son organisation sémantique.

2 CONTEXTE DE L'ÉTUDE

2.1 Le projet Emotirob

Le projet EmotiRob a pour but de concevoir un robot compagnon autonome (peluche animée) pour apporter du bien être à des enfants fragilisés, par exemple en cas de longue hospitalisation. Il fait suite à des expérimentations menées dans des centres de rééducation par le laboratoire Valoria de l'université de Bretagne Sud, avec un autre robot peluche : le phoque « Paro », lesquelles ont montré le réconfort que pouvaient apporter de tels robots compagnons (Saint-Aimé et al., 2007). L'objectif d'Emotirob est d'augmenter les capacités réactives du robot qui doit pouvoir simuler les six émotions primaires : joie, tristesse, dégoût, peur, surprise et colère, par des mouvements du corps, les traits du visage et l'émission de petits sons simples. De plus, ces différentes expressions émotionnelles doivent être une réaction adéquate aux propos tenus par l'enfant.

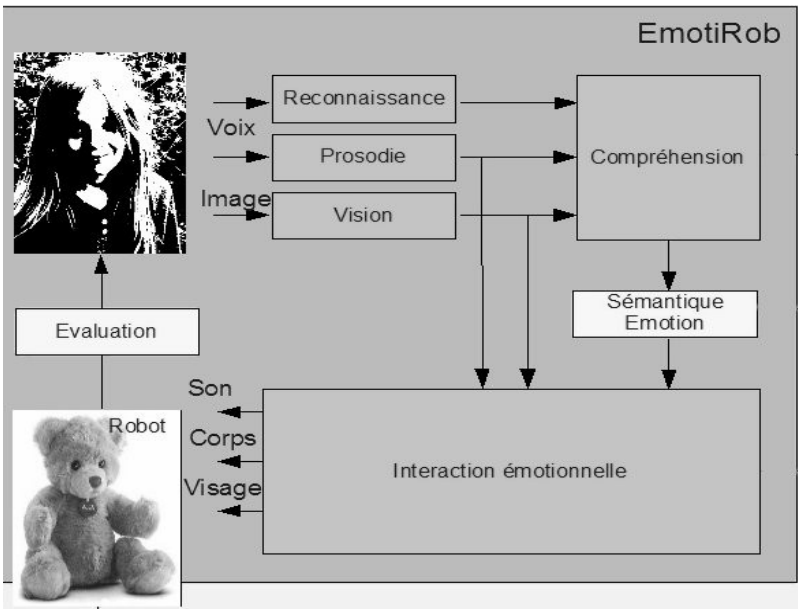


Figure 1 : Synoptique du projet EmotiRob

Le projet comporte deux parties distinctes : un module de compréhension qui interpréte les propos de l'enfant et un module d'interaction émotionnelle chargé de faire exprimer au robot une réponse émotionnelle appropriée. Le projet prévoit également une phase d'évaluation en situation réelle d'interaction dans un centre hospitalier pour mesurer la qualité des réponses et leurs impacts sur l'enfant. L'article concerne des études linguistiques qui ont été menées sur un corpus de contes, pour aider à l'implémentation du système de compréhension.

2.2 Le système de compréhension

Un système de reconnaissance vocale transcrit les paroles prononcées par l'enfant sous la forme d'une suite de mots. Les erreurs de cette reconnaissance et les spécificités de la parole spontanée imposent pour l'analyse l'utilisation de traitements robustes. Une solution classique mise en oeuvre dans des domaines d'application très étroits et pour des tâches très spécifiques consiste à définir des structures sémantiques figées. La compréhension peut alors être réduite à la détection de mots ou de séquences de mots clefs qui viennent remplir les différents paramètres de ces structures. De telles solutions ne sont guère envisageables dans le cadre de notre projet. En effet, l'une des difficultés majeures à laquelle nous nous trouvons confrontés est que l'on ne peut que formuler des hypothèses sur les propos que tiendra l'enfant à une telle peluche et qu'il est d'ailleurs probable que ces propos seront largement dépendants, et de l'enfant, et des capacités réactionnelles du robot qui sera mis entre ses mains. En tout état de cause, la compréhension doit couvrir un domaine relativement large. En revanche, l'âge des enfants concernés (5 ans environ) peut faire espérer que l'on peut se restreindre à un vocabulaire relativement limité.

Le module de compréhension choisi réutilise les principes de LOGUS (Villaneau et al., 2002), un système de compréhension conçu pour des domaines d'application relativement étendus. Plusieurs campagnes d'évaluation (Devillers et al., 2004) ont montré que ce système présentait une robustesse d'analyse satisfaisante sur de la parole spontanée. L'objectif de notre recherche est d'adapter ce système, initialement conçu dans le domaine test du renseignement touristique, au nouveau contexte applicatif du projet EMOTIROB. La tâche s'avère ambitieuse car, même si LOGUS a été conçu pour pouvoir fonctionner dans un domaine plus large que les domaines pour lesquels sont actuellement prévus les systèmes automatiques de compréhension de la parole, on a ici à gérer un

domaine d'application potentiellement illimité. Elle doit néanmoins être réalisable si l'on considère que le module de compréhension est autorisé à ne pas « tout comprendre ». Il est en effet prévu d'associer à l'interprétation des propos de l'enfant une étude de leur prosodie pour détecter ses états émotionnels.

2.3 Présentation générale du système LOGUS

Contrairement aux stratégies de compréhension adoptées dans la plupart des systèmes classiques, les structures sémantiques utilisées dans LOGUS ne sont pas figées. L'analyse doit donc être capable d'extraire de l'énoncé lui-même les relations entre les différents éléments qui le composent, en combinant les indices syntaxiques et sémantiques : les règles utilisées s'appuient à la fois sur la nature syntaxique des éléments et sur une connaissance sémantique du domaine de l'application, définie par un certain nombre de prédicats.

L'analyse se déroule suivant trois grandes étapes suivant le schéma représenté ci-dessous ; le lexique donne à chaque mot – ou groupe figé de mots -, une ou plusieurs définitions (lemmatisation). Ensuite, le chunking rattache les mots grammaticaux aux mots auxquels ils se rattachent, créant ainsi des petits groupes porteurs de sens. Enfin, la mise en relation de ces chunks aboutit à la construction de la formule logique qui représente le sens de l'énoncé.

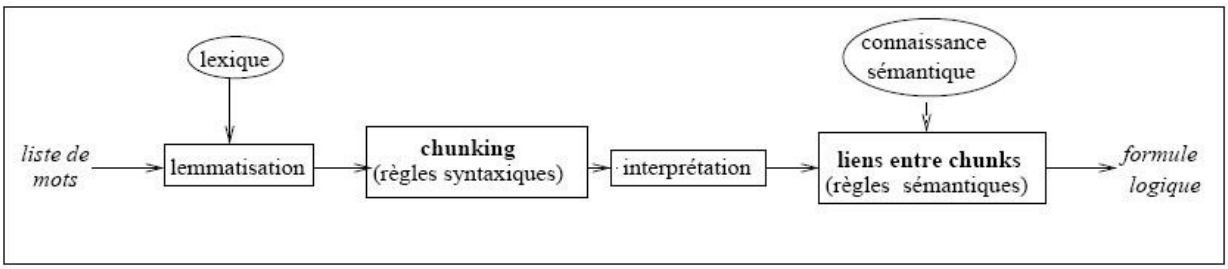


Figure 2 : Les différentes étapes dans Logus

Chacune de ces étapes nécessite une adaptation plus ou moins lourde pour réaliser la compréhension que nécessite le projet Emotirob.

2.3.1 Lemmatisation

Cette première étape consiste à remplacer chaque mot de l'énoncé par sa ou ses définitions, chacune d'entre elles correspondant à un triplet comprenant une étiquette syntaxique, une étiquette sémantique, et la représentation sémantique. L'adaptation de Logus au projet Emotirob nécessite de revoir presque entièrement le lexique pour ce qui concerne les mots non grammaticaux : noms, adjectifs et verbes d'action. Il est prévu également d'ajouter une étiquette en rapport avec l'émotion sur chacun des mots.

Pour définir le vocabulaire à prendre en compte, nous avons récupéré des informations sur le langage des enfants à partir de données disponibles que sont Novlex (Lambert, Chesnet, 2001), Manulex (Lété et al., 2004) et une étude de Dominique Bassano.

Manulex est une base de données lexicale qui fournit les fréquences d'occurrences de 23.900 lemmes et 48.900 formes orthographiques extraits d'un corpus de 54 manuels scolaires de lecture. Librement accessible sur internet, la base est à la disposition des chercheurs qui travaillent notamment sur l'acquisition de la lecture. La base de données lexicales Novlex est un outil permettant d'estimer l'étendue et la fréquence lexicale du vocabulaire écrit adressé à des élèves francophones de l'enseignement primaire. Elle a été constituée grâce à l'analyse de livres scolaires et extra-scolaires destinés à des élèves de CE2 (8-9 ans). Novlex est construit à partir d'un corpus d'à

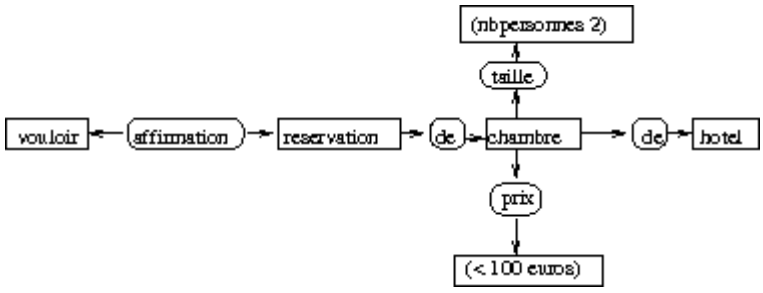
peu près 417 000 mots, ne comprenant ni noms propres, ni prénoms, ni noms de ville, ni onomatopées et ramenés en minuscule. De ce corpus, 20 600 entrées orthographiquement différentes (base d'occurrences) et 9300 racines lexicales (base lexicale) distinctes ont été extraites. La dernière source d'information utilisée est une étude menée par Dominique Bassano sur le développement du langage chez l'enfant. Cette étude reprend environ 1500 mots recueillis à partir de questionnaires donnés à des parents d'enfants. On trouve dans cette étude les verbes que connaissent les enfants, les adjectifs et les noms propres classés en différentes catégories telles que les animaux, la maison, les repas, etc.

2.3.2 Chunking

Cette étape segmente l'énoncé en constituants minimaux (chunks) tout en analysant leur structure interne. Les chunks correspondent le plus souvent à des unités de sens qui représentent les objets de l'univers. La segmentation facilite donc la transition vers les traitements sémantico-pragmatiques ultérieurs. Mais surtout, elle est particulièrement bien adaptée à la langue orale spontanée. Il a en effet été démontré que ces constituants sont le lieu de réalisation privilégié des réparations à l'oral (Blanche-Benveniste, 2005). La portée limitée de la segmentation garantit donc une certaine robustesse tout en autorisant une analyse plus détaillée des énoncés oraux. Aucun élément n'est en effet ignoré à ce stade. Cette étape analyse la structure interne des constituants en plus de caractériser leurs frontières. Enfin, l'étiquetage et la segmentation reposent sur une connaissance syntaxique totalement indépendante de la tâche. Cette architecture présente donc d'indéniables atouts en matière de généricité. Les structures syntaxiques locales ne semblent pas présenter de différences importantes entre le langage des enfants et celui des adultes : l'étape de chunking n'a pas nécessité de modifications notoires du système.

2.3.3 Dépendances sémantiques

Cette étape conduit à la représentation sémantique finale du contenu propositionnel. Elle repose sur une connaissance des liens possibles entre différents objets présents dans l'énoncé. Cette connaissance se présente sous la forme de relations de dépendances prédicat/argument entre les différents objets du domaine. Par exemples, dans le système LOGUS initial, l'énoncé « *je veux réserver un hôtel pour deux personnes à moins de 100€ »* est divisé en cinq chunks *« [je voudrais] [réserver] [un hôtel] [pour deux personnes] [à moins de 100€] ».* Ensuite, les connaissances sémantiques permettent d'établir les liens sémantiques possibles entre ces chunks. Le système doit donc « savoir » qu'un hôtel est composé de chambres, qu'une chambre peut être réservée, et qu'elle peut avoir pour proriétés un nombre d'occupants et un tarif. La formule logique rendue par LOGUS peut être représentée par le graphe conceptuel représenté ci -dessous. Les concepts y figurent dans les boîtes rectangulaires et les relations conceptuelles dans les ellipses.



Définir l'ensemble de ces liens sémantiques est le plus gros travail d'adaptation à réaliser pour adapter Logus au projet Emotirob. Représenter l'ensemble des connaissances du monde est une chose impossible et nous n'avons pas l'ambition de créer un système capable de traiter de tous les domaines. Malgré cela et même en restreignant le domaine couvert à celui que connaît un jeune

enfant, le nombre d'objets et de concepts est beaucoup plus important et plus varié que dans l'application initiale. Il nous a donc semblé qu'une étude linguistique s'avérait nécessaire pour avoir une réelle connaissance des liens sémantiques effectivement utilisés dans le monde des enfants.

3 CORPUS, TRAITEMENTS ET MODÈLE D'ANALYSE

Le travail mené pour alimenter les connaissances du programme Logus a consisté à construire des patrons verbaux à partir d'une analyse de corpus. Il s'agissait d'abstraire du corpus des schémas verbaux associés à des catégories sémantiques qui permettraient au module de compréhension d'offrir une représentation sémantique à chaque énoncé. Pour ce faire, un corpus de test a tout d'abord été constitué, puis traité sémantiquement et enfin analysé manuellement.

3.1 Présentation du corpus

Le projet EmotiRob est principalement doté de deux corpus de texte. Le corpus Brassens regroupe des retranscriptions de prises de parole d'enfants dans un contexte scolaire. La tâche remplie par ces enfants est de faire la lecture à voix haute d'un conte préalablement inventé en classe. Ce corpus qui comporte environ 40 000 mots, fut principalement constitué dans l'objectif de tester la reconnaissance vocale, qui transcrit automatiquement les mots analysés par la suite par le module de compréhension. Des corpus oraux de ce type restent encore difficiles à obtenir pour le français, notamment si l'on adopte une analyse quantitative.

Les recherches se sont ainsi portées sur un corpus plus large, dont la constitution posait moins de difficulté. Il s'agit d'une collection de contes pour enfants, extraite automatiquement sur la toile, comportant environ 160 000 mots. Le corpus comprend 138 contes de longueur variable (entre 120 et 17 000 mots). Les histoires qu'il contient sont de nature diverse : la part est faite aux contes de fées classique, ainsi qu'aux contes d'animaux. Il contient également des histoires communes relatant le quotidien d'un enfant se rendant à l'école, mais certaines histoires sont aussi situées dans un contexte historiquement daté ou encore, décrivent l'arrivée d'extraterrestres sur terre.

Les types d'auteurs varient également : le site à partir duquel les contes ont été collectés apportait, pour la majorité, des informations quant au contexte de la création du conte. Par exemple, certans contes sont regroupés autour d'une thématique étudiée en classe et il était demandé aux enfants de broder une histoire autour de quelques éléments (à partir de personnages-types, par exemple, comme le tigre). Collecter ces informations et les associer aux fréquences de mots a permis de se forger une idée globale sur le type d'auteur et de constater leur répartitions :

Type d'auteur	Fréquence	Proportion
Conteur moderne adulte	63 217	39%
Enfant	53 109	34%
Inconnu	34 314	21%
Conteur classique	9 900	6%
Total	160 540	

Type d'auteur	nb de contes	Proportion
Enfant	70	51%
Inconnu	37	27%
Conteur moderne adulte	24	17%
Conteur clasique	7	5%
Total	138	

Tableau 1 : Auteurs des contes dans le corpus de contes pour enfants en termes de fréquence et de nombre de textes

On observe sur le tableau 1, que bien que le corpus comporte davantage d'histoires écrites par des enfants que par des adultes, les proportions en nombre de mots sont relativement similaires (39% et 34%); ceci s'explique par le fait que les enfants écrivent généralement des histoires courtes. À noter qu'un certain nombre de contes sont des adaptations de conte classique comme *Le Petit Chaperon Rouge* ou *Peau d'Âne*. Ces résultats doivent néanmoins être relativisés car il n'a pas été possible d'identifier la nature de la source pour moins d'un tiers des texte (environ 20% des mots du corpus;

étiquette *Inconnu*). Tous les textes sont écrits en français contemporain.

Le corpus semble relativement diversifié en termes de contenu ainsi qu'en terme de sources et cette variation interne peut être interprétée comme un gage de sa représentativité. En effet, en variant les sources, les phénomènes idiosyncrasiques propres à chaque auteur auront tendance à s'estomper pour ne collecter que des informations spécifiques au genre de texte. De plus, les contes partagent toutes un destinataire commun, les enfants : ce sont en majorité de courtes histoires destinées à être lues à des enfants afin de stimuler leur imagination.

3.2 Données de l'expérience

Afin de répondre aux besoins applicatifs, nous nous sommes focalisés sur le lexique de Logus. Le lexique fourni (construit dans les conditions décrites en 1) comporte, entre autres, 659 noms communs regroupés en catégories sémantiques générales et 235 verbes. La couverture offerte par le corpus est illustrée dans le tableau 2 : 76% des noms et 90% des verbes figurent dans le corpus de contes.

	Noms totaux	Verbes totaux	Noms en commun	Verbes en commun
Lexique Logus	654	234	494 (76%)	210 (90%)
Corpus de Contes	7291	1535	494 (7%)	210 (14%)

Tableau 2 : Noms et verbes communs au lexique Logus et au corpus de contes

Il faut noter que les chiffres des verbes et noms totaux du corpus de contes doivent être corrigés car ce calcul a été effectué sur la base des étiquettes morpho-syntaxiques fournies par l'analyseur automatique Tree-tagger, et comporte ainsi de nombreuses erreurs d'étiquetage; ils sont néanmoins indiqués à titre indicatif. Les noms du lexique Logus qui n'ont pas été retrouvés en corpus sont des formes plurielles, des abbréviations ou encore des noms désignant des entités dont il n'était pas question dans le corpus (par exemple, « kangourou » ou « kiwi »). L'expérience n'a été menée que sur les verbes et noms présents dans le lexique Logus, soit des données vérifiées.

Parmi les 210 verbes en commun, 90 verbes ont été sélectionnés pour l'analyse. Il s'agit de verbes dont la fréquence dans le corpus de contes dépasse 30 occurrences. Certains verbes (10 en tout) ont été écartés de l'analyse, comme les modaux et auxiliaires ainsi que quelques verbes comme « arriver » ou « voir » que les délais de l'expérience n'ont pas permis d'analyser. Les fréquences des 90 verbes sont rappelées dans le tableau 3 :

Verbes	Fréquence	Verbes	Fréquence	Verbes	Fréquence	Verbes	Fréquence	Verbes	Fréquence
dire	731	commencer	121	habiter	73	écouter	49	se promener	34
savoir	290	crier	116	connaître	73	pousser	49	montrer	33
venir	286	s'appeler	114	comprendre	72	se coucher	49	se asseoir	33
prendre	265	attendre	113	retourner	70	boire	48	chanter	33
trouver	248	revenir	113	tirer	70	se réveiller	48	bouger	32
partir	224	appeler	111	finir	69	arrêter	48	couper	31
demander	216	parler	102	marcher	69	tenir	46	sourire	30
falloir	202	croire	101	se approcher	69	réussir	44	se reposer	28
donner	191	aimer	101	jouer	64	avancer	42	éteindre	18
regarder	187	laisser	95	se lever	62	pleurer	42	battre	17
passer	180	courir	95	dormir	61	rire	41		
entendre	177	aider	95	préparer	59	conduire	41		
sortir	168	rentrer	90	cacher	58	jeter	40		
manger	158	ouvrir	87	perdre	54	emmener	40		
rester	149	monter	87	porter	54	lancer	39		
chercher	140	raconter	86	apporter	54	attraper	39		
décider	135	sauter	80	descendre	51	sentir	38		
devenir	129	retrouver	79	rendre	50	tuer	36		
penser	123	essayer	77	poser	50	servir	35		
tomber	121	voler	73	oublier	50	tourner	35		

Tableau 3 : Fréquence dans le corpus de contes des verbes retenus pour l'expérience

8527 occurrences verbales ont en tout été examinées. Chaque verbe a été analysé en fonction de son environnement grammatical et leurs patrons grammaticaux ont pu être identifiés. Le tableau 4 expose les patrons grammaticaux du verbe « appeler » de fréquence 111 (associée à la forme « s'appeler » de fréquence 114), présentés dans le format CPA (cf. 2.3).

Proportion	**Pattern** / *Implicature*
51%	**[[Type]] s'appeler {[N]}**
	The name of [[Type]] is [N]
21%	**[[Type 1]] appeler [[Type 2]] ({en souvenir de [[Type 3]]})**
	[[Type 1]] invents a name [N], which may take its origin from [[Type 3]] to refer to [[Type 2]]
21%	**[[Type 1]] appeler [[Type 2]] ([[Speech Act]])**
	[[Type 1]] requires attention from [[Type 2]] by pronouncing his name or something else in a [[Speech Act]]
5%	**[[Type 1]] appeler [[Type 2]]**
	[[Type 1]] makes a telephone call in order to talk to [[Type 2]]
1%	**[[Type]] appeler {au secours | à l'aide}**
	[[Type]] shouts for help
<1%	**[[Type 1]] appeler [[Type 2]] {à [[Type 3]]}**
	[[Type 2]] is asked by [[Type 1]] to attend an [[Event]] specified in the prepositional phrase

Tableau 4 : Patrons grammaticaux du verbe « appeler » identifiés dans le corpus de contes

Dans le corpus de contes, ce verbe est essentiellement employé pour nommer des entités (deux premiers patrons dans le tableau 4). Un autre usage (troisième patron) correspond au sens « attirer l'attention de quelqu'un ». Cet usage est proche de celui décrit dans le quatrième patron, à la différence près que l'appel est de nature téléphonique, c'est-à-dire que l'entité en position sujet utilise un appareil de communication (radio, talkie-walkie, téléphone). Un usage mineur de ce verbe a également été isolé dans un patron car il correspond aux formules figées « appeler au secours » et « appeler à l'aide », dans lesquels les noms « aide » et « secours » désignent le but de l'appel.

Les noms, quant à eux, sont regroupés en classes sémantiques et portent des étiquettes évoquant au mieux la nature sémantique de ses membres, comme les classes [boisson], [moment], [vehicule]. Les noms du lexique retenus pour l'analyse sont ceux qui entraient dans les patrons d'un des 90 verbes, au nombre de 397, formant un ensemble de 4397 couples verbo-nominaux. Le tableau 5 présente les classes les plus fréquentes (54 classes sont concernées au total pour ces relations verbo-nominales), associées à leur productivité (nombre d'éléments différents d'une classe donnée). 8768 arguments des 90 verbes sont des noms (un verbe peut avoir plusieurs sujets ou objets, ce qui explique que ce chiffre dépasse celui des occurrences verbales), sachant que plus d'un tiers correspond à la classe *Humain* (34%) et moins d'un quart à la classe *Animaux* (20,4%), ce qui représente ensemble plus de la moitié des occurrences des noms (54,6%). La classe *divers* regroupe des noms non encore classés.

Type	Fréquence	Productivité	Exemples
humain	3004	18	enfant(1506) fille(902) garçon(255)
animaux	1789	46	chat(207) cheval(140) tigre(139)
divers	967	102	maison(73) histoire(55) arbre(47)
humain_fct	497	16	princesse(169) roi(124) clown(44)
humain_contes	486	5	lutin(204) fée(129) sorcière(111)
anim_génér	284	3	animal(191) oiseau(90) bête(3)
partie_corps	215	26	tête(32) main(30) pied(22)
partie_pieces	112	11	porte(45) fenêtre(18) pièce(15)
ciel	104	5	soleil(32) lune(30) ciel(19)
boisson	104	6	eau(64) lait(27) café(10)
fleur	103	2	fleur(78) rose(25)
jour/nuit	100	2	jour(55) nuit(45)
jeux	85	9	jouet(45) poupée(19) ballon(10)
anim_spéci	79	2	dragon(53) monstre(26)
vehicule	71	9	voiture(24) camion(17) bateau(9)
moment	68	3	heure(38) matin(16) journée(14)
pluie/neige	58	3	vent(43) pluie(10) neige(5)
meubles	44	7	table(17) chaise(13) coffre(5)
moment repas	42	7	fromage(11) petit-déjeuner(8) déjeuner(6)
lieu ville	41	6	travail(14) ferme(8) magasin(8)
repas_princi	39	9	poulet(17) soupe(8) galette(5)
feu	35	2	feu(33) fumée(2)
pieces	33	6	chambre(19) bureau(4) couloir(3)
vet_accessoire	33	5	sac(16) bague(9) valise(5)
outils	30	6	allumette(11) clé(7) balai(5)
fruits	29	7	fraise(8) banane(6) noisette(6)
vet_pied	28	3	chaussure(20) chaussette(7) chausson(1)

Tableau 5 : Fréquence, Productivité et exemple des classes sémantiques du corpus de contes

3.3 Modèle sémantique appliqué

Le modèle sémantique appliqué au corpus est *Corpus Pattern Analysis*, proposé par Hanks (CPA; Hanks, 2008). Les travaux de ce lexicographe se situent dans le cadre de la linguistique de corpus (Sinclair, 1991), discipline qui pose le texte comme base à toute analyse. Hanks s'inscrit dans une perspective d'analyse guidée par le corpus (*corpus-driven*) et adopte par là le principe du

contextualisme du sens : le sens d'un mot se définit selon ses usages (*meaning is use*; cf. Firth 1957) et définir un mot revient à définir son environnement linguistique. L'étude quantitative de ce contexte linguistique permet d'identifier les patrons majeurs associés à une forme donnée. Dans cette perspective, le projet CPA (un site internet permet de consulter les patrons créés; cf. *http://nlp.fi.muni.cz/projekty/cpa/*) vise à construire un dictionnaire des principaux patrons des principaux verbes de l'anglais (« all the normal patterns for all the normal verbs in English »; Hanks & Ježek, 2008: 391) à partir de grands corpus suffisamment variés pour être représentatifs de ses usages courants, comme le BNC (British National Corpus; cf. *http://www.natcorp.ox.ac.uk/*). Hanks suggère de décomposer l'entrée d'un verbe en fonction de ses patrons habituels, ordonnés selon leur fréquence. Chaque patron contient les éléments nécessaires à son identification et chacun correspond en principe à un sens différent du verbe qui constitue son noyau.

Les patrons de CPA intègrent dans la mesure du possible, la nature sémantique de l'élément remplissant la fonction sujet, objet ou complément, ce qui est nommé le *type sémantique*. Ce type correspond à un noeud dans l'ontologie de référence (une version simplifiée de la *Brandeis Semantic Ontology* dans le cas de CPA; Pustejovsky et al. 2004: 56) et est censé couvrir toutes les unités pouvant apparaître dans une position du patron. Le type le plus général de l'ontologie est *[[Anything]]* que l'on retrouve par exemple utilisé dans le patron CPA suivant (tableau 6; cf. *http://nlp.fi.muni.cz/projekty/cpa/*) :

31%	**[[Anything]] be called [NOOBJ] {[N]}**
	The name of [[Anything]] is [N] *[[Anything]] may be an individual, a set of thing, a person, a human group, an idea, or anything*

Tableau 6 : Patron CPA le plus fréquemment associé au verbe « to call » en Anglais

Le tableau 6 fait figurer le patron le plus fréquent du verbe « to call » qui peut être traduit en français par « X s'appelle Y ». Dans ce cas, le type *[[Anything]]* indique que toute entité peut être sujet du verbe « to call » lorsqu'il est employé à la voix passive, autrement dit que toute entité peut être nommée et qu'il n'y a aucune restriction sémantique portant sur le sujet du verbe. On peut conclure que les unités observées en position sujet de cette forme verbale étaient très variées au point de ne pouvoir choisir d'autre type sémantique que *[[Anything]]*. Remarquons que ce patron n'exploite pas la hiérarchie des types sémantiques de l'ontologie et qu'il est équivalent à un patron grammatical tel que « [GN] be called [GN] » (où GN désigne tout type de groupe nominal). Ce cas est néanmoins particulier car la majorité des patrons CPA manipulent d'autres types sémantiques plus spécifiques. Par exemple, le verbe « to cry » (*pleurer*) est associé au patron *[[Human]] cry [NO OBJ]* dans 57% des énoncés analysés par Hanks (sur un échantillon de 1113 occurrences) et fait référence à un humain qui pleure comme dans l'exemple (1) :

(1) As each person began to **cry**, a rebirther helped to calm them down. [A48]

Cependant, la liste d'unités lexicales observées dans une position donnée ne correspond pas toujours à un seul type sémantique. Hanks parle dans ce cas d'alternance de type (alternation of semantic type; cf. Hanks & Ježek, 2008 : 397) et deux solutions sont possibles : (1) différencier les patrons en fonction de cette alternance ou (2) intégrer cette alternance au sein du patron.

Un patron est généralement créé lorsqu'un glissement sémantique peut être identifié. Deux patrons peuvent ne varier par exemple qu'en fonction du type sémantique en position *objet*. L'exemple proposé par Hanks est « to fire a person » (*virer quelqu'un*) et « to fire a bullet » (*tirer une balle*), dont la variation du type (*[[Human]]* contre *[[Projectile]]*) corrèle avec la variation de sens du patron.

Il est également possible qu'une alternance de type intervienne mais n'entraine pas de variation sémantique du patron. Pour rendre compte de cette alternance de types, il faut s'assurer qu'il n'existe

pas un type sémantique plus abstrait mais plus approprié qui convienne pour décrire la totalité des unités lexicales observées dans une position donnée. Si tel n'est pas le cas, Hanks propose tout de même d'inscrire cette alternance dans le patron en spécifiant dans la position concernée, la liste des types sémantiques possibles. Le type le plus fréquemment rencontré apparaîtra en premier et l'autre (ou les autres) constituera un type alternatif également autorisé dans le patron. Les alternances de type peuvent s'expliquer en contexte par des phénomènes métonymiques, dans lesquels une chose est désignée par la dénomination d'une autre avec laquelle elle entretient une relation sémantique, comme l'individu par l'organisation dont il fait partie (*Matignon* pour *premier ministre*) ou le contenu par son contenant (*verre* pour *eau*). Ces relations ne donnent néanmoins pas lieu à la création d'un nouveau patron :

> 1. Les alternances sont liées au patron CPA principal à travers les mêmes mécanismes de modification de sens que ceux qui permettent d'interpréter les coercions. Néanmoins, les alternances constituent des réalisations différentes de la même norme.[i] [Pustejovsky et al, 2004 : 55]

À travers cette citation, Hanks rappelle que c'est le sens du patron qui prime sur les variations de sens de ses éléments. Lorsqu'une alternance de type n'occasionne pas de variation sémantique du patron, elle est intégrée dans le patron.

3.4 Alternances de type et Coercion

CPA ne modélise pas les relations sémantiques qu'entretiennent des types sémantiques s'alternant dans une même position. Hanks fait néanmoins régulièrement référence au modèle du Lexique Génératif (GL par la suite; Pustejosky, 1995), pour proposer des pistes de recherche. Il aborde notamment le mécanisme de coercion (Pustejovsky, 1995: 111), désignant un mode de composition sémantique particulier résultant d'un conflit entre le type attendu par un prédicat et le(s) type(s) observé(s). La définition proposée par Hanks et Ježek est la suivante :

> La coercion de type est une opération d'ajustement de type qui se produit lorsqu'aucune des préférences selectionnelles d'un prédicateur n'est satisfaite par le type d'un nom avec lequel il se combine dans un texte particulier. Dans ce cas, la coercion de type est invoquée pour expliquer comment une combinaison verbe-argument non satisfaite peut être interprétée. [Hanks & Ježek, 2008 : 395][ii]

Ces derniers donnent un ensemble d'exemples tirés de corpus anglais, dont l'alternance [Location]/ [Event] (*lieu / événement*), observée dans la liste des objets du verbe « to attend » (Hanks & Ježek, 2008 : 394) :

> attend
> Direct Object:
> a. [Event]: meeting, wedding, funeral, mass, game, ball, event, service, premiere
> b. [Location]: clinic, hospital, school, church, chapel
> “About thirty-five close friends and relatives attended the wedding”.

---

i Alternations are linked to the main CPA pattern through the same sense-modifying mechanisms as those that allow for coercions to be understood. However, alternations are different realizations of the same norm. [Pustejovsky et al, 2004 : 55]

ii Type coercion is an operation of type adjustment that occurs when none of the selectional preferences of a predicator match the type of a noun that it combines with in a particular text. In this case, type coercion is invoked to explain how a mismatching verb-argument combination can be interpreted. [Hanks & Ježek, 2008 : 395]

“For this investigation the patient must attend the clinic in the early morning”.
“He no longer attends the church”.

Les lieux apparaissant en position objet (*church*, *clinic*) sont alors réinterprétés comme les événements qui s'y produisent. Il propose donc d'intégrer cette alternance dans le même patron (2) :

(2) [[Human]] attend ([[Event | {Location = Functional}]]) [http://nlp.fi.muni.cz/projekty/cpa/]

Le modèle GL propose des opérations sémantiques qui formalisent ce genre de cas. Une première différence avec CPA est que GL focalise sur les noms, alors que CPA a été principalement appliqué aux verbes. Une seconde différence est que les prédicats du Lexique Génératif n'autorisent pas, dans la structure lexicale des verbes à proprement parler, d'alternance de type, alors qu'elles sont partie intégrante des patrons dans CPA. Pour expliquer les alternances de type, Pustejovsky propose d'enrichir la structure lexicale des noms. Voici trois exemples de formalisation sémantique rendant compte des alternances de type en contexte :

- La structure *Qualia* (Pustejovsky & Ježek, 2008 : 4) renferme des informations comme la finalité (*télique*) de l'objet désigné par l'unité lexicale. Par exemple, le télique d'un verre est de porter du liquide, ce qui autorise sa composition avec des verbes comme *boire* qui attendent un liquide en objet.
- Le *type formel* (équivalent au type dans l'ontologie) d'une unité lexicale peut être un objet complexe (*dot object*) comme « livre » (Pustejovsky & Ježek, 2008 : 11), dont la structure lexicale retient son sens d'*objet* (soulever un livre) et d'*information* (lire un livre).
- Les *attributs conventionnels* (*Conventionalized Attributes*) récemment ajoutés (Pustejovsky & Ježek, 2008 : 23) désignent des propriétés typiques d'entités, comme celle d'un oiseau de chanter.

Ainsi, la coercion consistera principalement à identifier dans la structure lexicale d'un argument, l'information qui autorisera la composition avec le type attendu d'un prédicat.

Nous nous fondons essentiellement sur le modèle CPA pour l'analyse de corpus, mais notre objectif est d'enrichir l'ontologie de Logus. Plutôt que de proposer une refonte de l'ontologie, nous avons opté pour l'ajout de liens entre types à travers l'intégration de deux opérations qui pourraient être assimilées aux mécanismes de GL, il s'agit des *propriétés saillantes* et des *conversions de type*. La connaissance sémantique de Logus ne prévoit ni de type complexe, ni de structure Qualia dans l'usage qu'en fait Pustejovsky. Néanmoins, enrichir la structure des unités lexicales des noms (et par conséquent de l'ontologie) permettrait à Logus d'associer également les types alternatifs encodés dans les patrons verbaux. Les propriétés saillantes sont inspirées des attributs conventionnels de GL (Pustejovsky & Ježek, 2008), sont associés uniquement à des unités lexicales (non à des types sémantiques) et désignent un type sémantique possible alternatif pour cette unité. Le problème se pose par exemple lorsque l'unité « concert » de type *événement*, est objet du verbe entendre (3) :

(3) Il se dé-rhume pour attirer l'attention du malfaiteur et se concentre pour se rappeler l'ennui du dernier *concert* de Beethoven qu' il a entendu

Pour prendre en compte cet exemple, l'unité « concert » sera associée à la propriété saillante *son* : un concert est un événement dont la particularité est d'être « sonore », soit un type particulier d'événement, un « événement sonore ». Les propriétés saillantes permettent ainsi de distinguer des sous-types sémantiques, comme celle des « instruments de musique » classés sous le type *Objets*, dont une des propriétés saillantes est de produire des sons comme l'attestent les exemples (4) et (5) :

(4) Elle entend la *trompe* de Monsieur Seguin mais ne veut pas retourner à la ferme

(5) Je me demandais pourquoi ils s' agenouillaient tous quand j'entendis un *clairon*

Ces sous-types représentent l'organisation sémantique des unités lexicales dans le corpus et contexualisent en quelques sortes l'ontologie de référence. L'objectif est de contextualiser les connaissances sémantiques et de proposer une version « naturelle », c'est-à-dire « fondée sur corpus » et respectueuse des données, de l'ontologie (El Maarouf, à paraître). Le fait d'externaliser ces nouvelles informations sémantiques de l'ontologie, permet, d'une part, de préserver la structure hiérarchique de l'ontologie qui demeure utile pour la plupart (cf. 3.1), et d'autre part d'envisager l'association d'unités lexicales et des types sémantiques en fonction d'un genre ou domaine de texte (cf 3.2 et 3.3).

Dans le cas où de nombreux éléments appartenant à une même classe possèdent des propriétés saillantes, nous avons prévu la possibilité d'associer deux classes par le mécanisme de *conversion de type*. Ce n'est pas le cas de *événement*/*son*, car peu d'événements sont sonores, mais c'est en revanche la cas du type *humain* qui apparaît souvent en tant qu'objet du verbe « entendre » (6) :

(6) Tout à coup, il *entend* Mélisa l'appeler de la cuisine

Dans notre corpus, tous les humains produisent des sons, mais il aurait tout à fait pu en être autrement, s'il y avait été question de personnes muettes par exemple.

4 ALTERNANCES DE TYPES ET ENRICHISSEMENT DE L'ONTOLOGIE

Cette partie décrit les différents phénomènes d'alternance rencontrés en corpus et le choix retenu pour enrichir l'ontologie. Nous décrivons d'abord l'exemple du verbe « s'approcher » comme exemple de notre démarche. Puis nous décrirons les conversions de type et les propriétés saillantes identifiées avant d'aborder une cause d'alternance particulière : les formules idiomatiques.

4.1 Analyse des alternances du verbe « s'approcher »

Le verbe « s'approcher » apparaît 69 fois dans le corpus, 29 fois dans le patron *[SUJ V de]* et 26 fois sur le mode intransitif *[SUJ V]*.

Patron	Fréquence
se approcher [SUJ : humain]	16
se approcher [SUJ : animaux]	10
se approcher [SUJ : divers]	3
se approcher [SUJ : animaux] [de : animaux]	3
se approcher [SUJ : humain] [de : divers]	3
se approcher [SUJ : humain_contes]	3
se approcher [de : divers]	3
se approcher [SUJ : animaux] [de : lieu ville]	2
se approcher [du : humain_fct]	1
se approcher [SUJ : anim_génér]	1
se approcher [de : humain_contes]	1
se approcher [SUJ : ciel]	1
se approcher [de : appareils]	1
se approcher [SUJ : humain_fct]	1
se approcher [SUJ : humain_fct] [de : divers]	1
se approcher [SUJ : humain] [de : animaux]	1
se approcher [SUJ : humain_fct] [de : humain]	1

Tableau 7 : Fréquence de patrons sémantiques de surface du verbe « s'approcher » du lexique Logus

Les autres patrons relevés sont dérivés de ces deux patrons principaux en y ajoutant tantôt une relation de manière (7) ou de but (8).

(7) Le minet s'approche *tout doucement* et se faufile dans le feuillage à pas de souris .
(8) Il s'approcha *pour voir ce qu'il se passait*

Générer les différents patrons possibles de ce verbe permet d'observer la variation des types en fonction des relations grammaticales retenues. Le tableau 7 donne les fréquences des patrons sémantiques possibles du verbe « s'approcher ». Tous les patrons possibles de ce verbe ne sont pas ici retenus car ils peuvent mettre en jeu des adverbes, ou des syntagmes qui ne figurent pas dans le lexique Logus. Sont présentés dans le tableau uniquement les relations dont le nom figure dans ce lexique. À titre comparatif, le tableau 8 présente tous les patrons grammaticaux et les patrons sémantiques ayant été créés. Lorsque le nom ne figure pas dans le lexique, la relation est notée sans arguments. Par exemple, concernant le patron grammatical *se approcher [SUJ] [de]* (retrouvé 29 fois), 6 occurrences se combinent avec des mots absents du lexique en position *[SUJ]* et *[de]* et 5 occurrences se combinent avec un sujet de type *humain* mais un mot absent en position *[de]*.

Patron grammatical (PatGram)	Patron incluant les types (PatType)	Fréquence PatGram	Fréquence PatTyp
se approcher [SUJ] [de]	se approcher [SUJ] [de]	29	6
	se approcher [SUJ : humain] [de]		5
	se approcher [SUJ : humain] [de : divers]		3
	se approcher [SUJ : animaux] [de : animaux]		3
	se approcher [SUJ : animaux] [de : lieu ville]		2
	se approcher [SUJ] [de : divers]		2
	se approcher [SUJ : animaux] [de]		1
	se approcher [SUJ : humain_fct] [de : humain]		1
	se approcher [SUJ] [de : humain_fct]		1
	se approcher [SUJ] [de : appareils]		1
	se approcher [SUJ] [de : humain_contes]		1
	se approcher [SUJ : humain_fct] [de : divers]		1
	se approcher [SUJ : divers] [de]		1
	se approcher [SUJ : humain] [de : animaux]		1
se approcher [SUJ]	se approcher [SUJ : humain]	26	7
	se approcher [SUJ]		6
	se approcher [SUJ : animaux]		6
	se approcher [SUJ : divers]		2
	se approcher [SUJ : humain_contes]		2
	se approcher [SUJ : humain_fct]		1
	se approcher [SUJ : anim_génér]		1
	se approcher [SUJ : ciel]		1
se approcher [SUJ] [à]	se approcher [SUJ] [à]	26	1
se approcher [SUJ] [dans-l'-espoir-de] [d	se approcher [SUJ] [dans-l'-espoir-de] [de : diver	26	1
se approcher [MAN] [SUJ]	se approcher [MAN] [SUJ : animaux]	4	2
	se approcher [MAN] [SUJ : humain]		1
	se approcher [MAN] [SUJ]		1
se approcher [SUJ] [du-bord-de]	se approcher [SUJ : humain_contes] [du-bord-de	2	1
	se approcher [SUJ : animaux] [du-bord-de]		1
se approcher [SUJ] [pour]	se approcher [SUJ : humain] [pour]	2	1
	se approcher [SUJ] [pour]		1
se approcher [DEST] [MAN] [SUJ]	se approcher [DEST] [MAN] [SUJ : humain]	1	1
se approcher [DEST] [SUJ]	se approcher [DEST] [SUJ : humain]	1	1
se approcher [SUJ] [de] [pour]	se approcher [SUJ] [de] [pour]	1	1
se approcher [SUJ] [près-de]	se approcher [SUJ] [près-de]	1	1

Tableau 8 : Fréquence des patrons grammaticaux (PatGram) et sémantiques de surface (PatType)

Les patrons complets et croisés avec l'ontologie (tableau 7) ne représentent qu'environ 50% (31) des occurrences totales des patrons grammaticaux extraits du corpus (tableau 8). Ces 13 patrons « sémantiques » correspondent en fait à deux patrons grammaticaux cités plus haut (*X s'approcher* et *X s'approcher de Y*) et se distinguent par les types sémantiques trouvés. Concernant le patron *X s'approcher*, on retrouve essentiellement des noms d'animés (sauf pour *divers* et *ciel* sur lesquels nous reviendrons). En effet, dans l'ontologie, les classes *humain*, *humain_contes* et *humain_fct* sont regroupées sous la classe *Humain* elle-même dépendante hiérachiquement de *Animés*. De même, *anim_génér* et *animaux* sont classés sous *Animaux*, qui est également de type *Animés*. La classe la

plus appropriée pour les sujets de ce patron semble donc être *Animés*, d'autant que cette alternance ne provoque pas de changement de sens du patron.

Le membre de la classe *ciel* est « lune » illustré par l'exemple (9) (où la lune est également sujet du verbe « remarquer ») :

(9) La lune remarqua cette courageuse petite fille qui avançait difficilement dans l'obscurité. Doucement elle *s'approcha* et éclaira son chemin, comme pour l'encourager .

Ce cas n'est pas isolé dans la classe *ciel* : nous avons constaté une forte tendance de ses membres à être employés comme des animés. Pour en rendre compte dans l'ontologie, nous proposons donc de créer un lien entre *ciel* et *Animés*.

Les membres de la classe *divers* sont « bruit » et « plante » qui n'avaient pas encore été classés dans l'ontologie à l'heure de l'expérience. Voici leurs occurrences associées :

(10) Ils entendent un bruit énorme s'approcher, comme un tremblement de terre.
(11) Alors qu'il poursuivait son chemin, il vit des plantes qui lui semblaient s'approcher.

L'exemple (11) nous enseigne que dans les contes, les plantes peuvent se mouvoir. Nous avons créé une classe temporaire *Végétaux* qui constituerait son type principal et lié uniquement l'unité lexicale à la classe *Animés* par propriété saillante. En revanche, le sens du patron en (10) varie, car lorsqu'un son s'approche, cela implique certes qu'une entité s'approche en produisant du son mais signifie que le son devient de plus en plus audible. Nous interprétons cela comme un patron différent, où le sens serait « devenir de plus en plus audible » dans lequel le sujet serait de type *son* par exemple.

Hormis les exemples de *ciel* et de *Végétaux*, la classe *Animés* est donc choisie pour représenter la majorité des occurrences en position sujet dans le patron *X s'approcher*. Les patrons rencontrés sont illustrés en (12) :

(12) 1. [SUJ=Animés] s'approcher
2. [SUJ=Son] s'approcher

En ce qui concerne le patron *X s'approcher de Y*, on observe le même type de variation pour la position sujet. En revanche, les noms régis par la préposition « de » varient d'une autre manière. On y retrouve des objets, des lieux et des animés. Pour autant, le sens du patron ne change pas car il s'agit systématiquement du sens « s'approcher d'un point de référence concret ». L'analyse des occurrences montre que la propriété d'animé, pour la position complément, n'est pas essentielle pour ce verbe : les entités ne sont pas considérés en tant qu'elles sont capables de se mouvoir de façon autonome, mais seulement en tant qu'elles possèdent un corps ayant une concrétude, positionné dans un espace, dont on peut s'approcher.

Cette propriété de concrétude est partagée par toutes les unités que l'on retrouve dans cette position verbale -si tant est que l'on accepte de considérer que les êtres extraordinaires comme les sorcières existent et possèdent un corps dans les contes de fées. Nous avons donc retenu le choix de créer une classe *Concret* et y avons lié les types *Objets*, *Lieux* et *Animés* par le mécanisme de conversion de type. Une occurrence n'entre pas dans ces catégories :

(13) Sa maîtresse l'a trop souvent grondée quand elle osait s'approcher de ses *travaux*.

Dans l'exemple (13), il s'agit d'une chienne qui s'approche des tableaux d'arts de sa maîtresse. Dans l'ontologie, « travail » est classé comme une activité, mais désigne ici en fait le résultat de cette

activité, un objet concret qui sera de type *Concret*. Nous avons choisi de créer une propriété saillante pour « travail » renvoyant au type *Concret*. Il est donc possible de proposer le patron en (14) pour rendre compte de toutes ces occurrences :

(14) [SUJ=Animés] s'approcher [de=Concret]

La majorité des alternances apparentes des types sémantiques en position d'argument vis-à-vis du verbe « s'approcher » ont ainsi pu être accomodées en choisissant le niveau d'abstraction approprié dans l'ontologie. Trois patrons ont pu être distingués, dont l'un fondé sur une alternance de type (*Animés*/*son*).

4.2 Les conversions de type

Il ressort de l'expérience que les verbes retenus montrent des alternances difficilement compatibles avec la structure hiérarchique de l'ontologie. Pour nombre de cas, il a donc été jugé utile d'enrichir la structure des unités lexicales. Nous nommons *propriétés saillantes* l'ajout d'un type alternatif à une unité lexicale et parlons de *conversions de type* lorsque toute la classe est concernée par cette propriété. Comme il est possible que ces associations sémantiques ne se révèlent que vis à vis de certains verbes et que nous cherchons à éviter la surgénéralisation, les propriétés saillantes et les conversions de type n'ont été associées, dans un premier temps, qu'au verbe en question et dans une position donnée du patron.

La multiplicité des types auxquels peut appartenir une même unité lexicale observée en corpus met en relief le fait qu'une unité appartient à plusieurs classes et que seules certaines facettes « s'activent » lorsqu'elles sont combinées avec des verbes. Une unité lexicale hérite ainsi des propriétés de classes différentes. Un membre de la classe *Humain*, peut ainsi être considéré en tant qu'*être rationnel*, lorsqu'il est combiné avec des verbes cognitifs, un *locuteur* lorsqu'il se combine avec des verbes de parole ou un *corps* lorsqu'il est objet de verbes matériels comme « frapper », ou encore comme *lieu*, lorsqu'il est régi par la préposition « sur » comme dans « tomber sur » (15).

(15) Un ouistiti très malin et très farceur décroche la liane, Madeleine tombe par terre sur le pauvre Masilo

L'avantage de traiter séparément les mécanismes d'alternance sémantique est de pouvoir ne pas les utiliser dans un autre contexte et ainsi proposer des opérations de conversions de type ou des propriétés saillantes d'unités lexicales en fonction des corpus. À titre d'exemple, les sujets du verbe « dire » illustrent un phénomène d'alternance de type sémantique propre aux contes (tableau 9):

Type	Frequency	Proportion
Humain	544	75%
Animaux	107	15%
autres	71	10%

Tableau 9 : Alternance de type sémantique en position sujet du verbe « dire »

On observe que si, comme on pouvait s'y attendre, les humains sont le plus souvent sujet du verbe « dire », une part non négligeable des sujets sont des animaux (15%) et 10% des sujets sont de type autre (notamment des *Objets* et des *Végétaux*). Ce phénomène s'explique par la personnification massive des personnages dans les corpus de contes, dans lesquels les animaux, par exemple, sont dotés de caractéristiques « habituellement » attribués aux êtres humains, comme la capacité à parler. Il ne conviendrait néanmoins pas dans ce cas de regrouper les entités sous le type *Animés* car il semble s'agir plutôt d'une propriété du corpus qu'il vaut mieux externaliser.

Le verbe « dire » n'est pas unique en ce genre, cette alternance est principalement constatée pour les sujets de verbes de parole comme « raconter », « demander », « crier », « appeler », « parler » et de verbes de cognition comme « savoir », « comprendre », « oublier », « croire », « connaître », « décider ». Ces alternances ont été traitées comme des conversions vers le type *Humain*. Les principaux types convertis sont *Animaux* et *Jouet* et traduisent le fait que dans les contes, les végétaux, les jouets (17) et les animaux agissent, pensent et parlent comme des êtres humains :

(17) Il écoutait ce que *le petit garagiste* [Jouet] lui disait : « Bon, je vais essayer de travailler tout seul mais c'est dommage, *j'*aimais bien quand on était tout les deux... on a fait du bon boulot avec la voiture rouge

Le tableau 10 présente les principales conversions de type rencontrées dans le corpus de contes :

Type converti	Type de destination
Végétaux Jouet	Animés
Lieux_ville	Bâtiment
Moment	Événement
Végétaux Animaux Jouet	Humain
Véhicule Vaisselle Végétaux Bâtiment Contenant Humain Animaux Animés	Lieu
Vaisselle Animaux	Nourriture
Véhicule Bâtiment	Objet_concret
Animé	Son

Tableau 10 : Principales conversions de type dans le corpus de contes

On notera le potentiel d'une redondance lorsque l'on compare les conversions vers *Animés* et vers *Humain*, car il pourrait être considéré que si un type est déjà converti en *Humain*, il ne serait pas nécessaire de spécifier qu'il devrait également être converti en *Animés*, qui s'hériterait par la structure hiérarchique de l'ontologie. Néanmoins, différents verbes nécessitent différentes conversions : certains verbes nécessitent la conversion [*Végétaux->Animés*] par exemple parce qu'ils n'acceptent que des Animés (notamment des verbes de mouvement) comme sujet mais d'autres verbes nécessitent la conversion [*Végétal->Humain*] (comme les verbes de parole et les verbes de cognition). De plus, le processus consistant à rendre une entité animée est un processus différent de la personification, consistant à rendre une entité humaine et il convient de les distinguer, ne serait-ce que pour limiter les surgénérations possibles. Notons enfin que les animaux n'ont pas à être convertis en *Animés* puisqu'ils héritent de cette propriété par l'ontologie.

On observe dans le tableau 10 que la conversion vers le type *Lieu* est celle qui donne lieu à la plus grande diversité quant au type converti. Il s'agit de patrons dont l'une des positions attendues est le *lieu* comme en (18) mais où l'on rencontre également des *Humain* (19) ou des *Animaux* (20).

(18) Etant donné que les brigands voyaient à l'avance qui s'approchait de leur *cabane*, le prince disparut à l' aide de son épée
(19) Le premier qui s'approcherait de cet *homme* crèverait la bulle
(20) Elle s'approche du *chat* qui dort lui aussi

Enfin une conversion de type recensée dans le tableau 10 rend compte de la capacité d'*Animés* à produire des sons et à être directement objet de verbes comme « entendre » :

(21) Mais déjà, dehors, on entendles *trolls* qui reviennent de leur promenade!
(22) Elle entend le grand *sorcier* qui ronfle

4.3 Les propriétés saillantes

Certaines alternances ne concernent que trop peu de membres d'une classe pour pouvoir être généralisées à des conversions de type. Dans ce cas, une propriété saillante émergeant des contextes est définie et est liée aux unités lexicales concernées. Les propriétés saillantes provoquent tantôt la création d'un nouveau patron et tantôt s'accomodent au type principal attendu par le patron. Nous avons réuni dans un tableau les propriétés saillantes (tableau 11) :

Propriété Saillante	Exemples	Nb
Lieu	bras, bureau, caillou, chaise, champignon, cheveu, ciel, eau, fil, four, herbe, lit, neige, pierre, table, terre, vague	17
Surface	arbre, bureau, coussin, dos, eau, épaule, genou, lit, marche, oreiller, poil, route, sac, table, tapis, toit	16
Contenant	assiette, boite, bouche, cage, coffre, gueule, livre, main, manche, noix, poche, sac, valise, verre	14
Volant	oiseau, monstre, sorcière, pigeon, papillon, enfant, dragon, canard, fusée, fée, animal	11
Humain	poupée, jambe, lune, vent, fruit, chaussure, arbre, feu, étoile, visage	10
Mobile	vent, nuage, roue, soleil, camion, ballon, étoile, mer	8
Chemin	route, rivière, rue, chemin, tunnel	5
Lumière	bougie, allumette, lumière, feu	4
Information	journée, peur, histoire, colère	4
Préhension	gueule, main, bras	3
Espace aérien	ciel, cour, nuage	3
Véhicule	chameau, cheval, balais	3
Visuel	tapis, fumée, arbre	3
Ouverture	barrière, fenêtre, porte	3
Animé	nuage	1
Boisson	verre	1
Obstacle	dragon	1

Tableau 11 : Principales propriétés saillantes et nombre d'unités lexicales concernés

Nous ne pouvons pas ici revenir sur chaque unité lexicale à laquelle a été attribuée une propriété saillante, mais nous donnerons quelques directions pour expliquer la manière dont ces propriétés ont été extraites du corpus.

La propriété saillante la plus employée est le *Lieu*. Certains meubles comme « bureau », « chaise » ou « lit » (23) sont en effet employés comme des lieux, mais ne sont pas classés en tant que tel dans l'ontologie; certains noms sont moins évidents à classer comme lieu (« ciel », « vague » (24)) et des unités comme « eau », « terre » (25) font référence à un lieu caractérisé par la matière dont il est constitué.

(23) Catastrophe, il tomba du *lit*.
(24) Pour se moquer de celui qui tombait dans les *vagues* monstrueuses, il eut juste le temps de crier
(25) Elle trouve les empreintes de Kakou dans la *terre* du chemin

Les *surfaces* sont un type d'objet proche des *lieux*, mais qui apparaît spécifiquement dans des positions de patrons où il est question de surface. Ces positions sont souvent introduites par la préposition « sur » comme dans les exemples (26) et (27) :

(26) Quoique étonnés par l'appétit inhabituel de leur père, les enfants sortirent tout ce qui restait et posèrent une nouvelle bouteille de Vodka sur la *table*
(27) Mélisa s'assoit sur le *sable* et réfléchit, soudain ... une idée lui vient en tête

La propriété de *contenant* est aussi une notion sémantiquement proche de celle de *lieu*, mais s'applique essentiellement à des objets ayant la particularité d'être clos (« coffre », « cage ») qu'il est possible d'ouvrir, dans lesquels on peut introduire d'autres objets ou entités, ou encore qui désignent des objets aux frontières délimitées (« assiette », « verre ») dans lesquels on peut placer des objets. Ces unités se combinent typiquement avec les verbes « ouvrir » et en complément de « sortir_de » (28), ou de « prendre dans » (29) comme dans les exemples suivants :

(28) La sorcière ouvre alors son *livre* de magie, sort de sa *poche* une étrange bouteille et se met à gronder
(29) La petite fille se leva et prit le lapin dans ses *mains*

La propriété *volant* s'applique à des entités douées de la capacité à se mouvoir de manière autonome dans les airs. Un nombre d'unités lexicales restreint (« oiseau », « pigeon », etc.) apparaît dans des patrons comme X battre des ailes (30), X se poser sur (31) ou encore X voler (32).

(30) C'est simple, lui expliqua Tireloui, *tu* arrêtes de battre des ailes et tu verras, peu à peu tu vas te mettre à descendre!
(31) Un *pigeon* se posa sur le rebord, prit dans son bec la lettre du petit enfant et s'envola dans les airs
(32) Le *dragon* leur indiqua le chemin en volant au-dessus d' eux

D'autres propriétés concernent la fonction de *préhension* qui distingue un sous-type de partie du corps destinées à saisir des objets et que l'on retrouve par exemple comme complément de « tenir dans » (33). Enfin, la propriété de *boisson* concerne uniquement « verre » qui traduit l'alternance bien connue de contenant-contenu en (34)

(33) Dans une *main* il tient une corde et de l'autre les rênes
(34) Du seuil il peut voir sa femme et son fils assis autour de la table, se pinçant le nez et buvant chacun un grand *verre* d' eau, à tour de rôle

4.4 Formules idiomatiques

Les alternances constatées en corpus peuvent être dues à des facteurs irréductibles à types sémantiques. Dans ce genre de cas, tel que le recommande Hanks il convient de créer un patron à part et d'inscrire dans ce dernier la forme et non le type de l'unité concernée. Les formules idiomatiques sont en partie responsables de la variation observée des types sémantiques.
Parmi, les sujets du verbe « dire », on trouve par exemple le nom « doigt » (35) :

(35) mon petit *doigt* me dit que...

Dans ce cas il s'agit d'une formule idiomatique, qu'il faut pouvoir isoler; il en est de même pour « nom » en (36) :

(36) ce *nom* ne me dit rien

Il s'agit ici d'un patron dont le sens diffère et qui peut être paraphrasé par le verbe « je n'ai jamais entendu ce nom ». Ces deux formules sont relativement figées et on peut choisir d'assigner le type sémantique de chacun de ces collocats (*nom* et *doigt*) en position sujet (comme en (37)) ou de spécifier dans le patron la forme exacte de l'unité lexicale (comme en (38))

(37) [SUJ=Information] [IOBJ=Humain] dire [OBJ={rien}]
[SUJ={nom}] [IOBJ=Humain] dire [OBJ={rien}]

(38) [SUJ=[[POSSESSIF]] [Partie_Corps]] [IOBJ=Humain] dire [que=[proposition]]
[SUJ=[[POSSESSIF]] {doigt}] [IOBJ=Humain] dire [que=[proposition]]

Afin d'éviter la surgénéralisation, nous avons choisi la solution présentée en (38). Le tableau 12

liste les patrons de quelques expressions idiomatiques extraites du corpus de contes et formalisées sur ce modèle :

Expression idiomatique	Patron correspondant
jouer un tour à quelqu'un	jouer [OBJ={un tour}] [IOBJ=Humain] [SUJ=Humain]
prendre son courage à deux mains	prendre_[OBJ={son courage}]_[à={deux mains}]
prendre ses jambes à son cou	prendre_[OBJ=[POSSESSIF] {jambes}]_[à=[POSSESSIF] {cou}]
tomber à genou	tomber_[à={genou}]
tomber à terre	tomber_[à={terre}]
lancer les bras au ciel	lancer_[OBJ={les bras}]_[au={ciel}]
ne pas en croire ses oreilles	ne croire pas [[[POSSESSIF]] {oreilles}]
le ventre crie famine	crier_[SUJ={ventre}]_[OBJ=famine]
faire venir l'eau à la bouche	faire venir_[SUJ={l'eau}]_[à={la bouche}]
rendre l'âme	rendre_[OBJ={l'âme}]_[SUJ=animé]
les larmes lui montent aux yeux	monter_[SUJ={larme}]_[à={yeux}]
tenir le coup	tenir_[OBJ={le coup}]
Le ciel leur tombe sur la tête	tomber_[SUJ={ciel}]_[IOBJ=humain]_[sur={tête}]

Tableau 12 : Formalisations d'expressions idiomatiques rencontrées en corpus

5 CONCLUSION

L'objet de cet article était de présenter un travail de recherche en cours de réalisation à l'université de Bretagne-Sud. Il s'agit de travaux transdisciplinaires entre linguistique et informatique, s'inscrivant dans le projet de robotique EmotiRob présenté en première partie. Il est encore difficile d'apprécier l'impact ou les retombées possibles d'une telle recherche avant la fin du projet et ce travail en appelle d'autres.

Rappelons qu'un des objectifs est d'évaluer l'état de la connaissance sémantique du module de compréhension, décrit en première partie, à travers une analyse de corpus. L'angle d'approche choisi pour éclairer la couverture du système est le phénomène de coercion, qui se traduit dans le texte par des alternances de types sémantiques dans des positions syntaxiques identiques. Cela implique tout d'abord la constitution d'un corpus, sa description et son analyse dans le cadre du modèle sémantique choisi : *Corpus Pattern Analysis* (seconde partie de l'article). Le corpus de test choisi n'est pas nécessairement à l'image de la situation dans laquelle se retrouvera le robot *in fine*, mais il permet d'élaborer une première expérience de corpus sur la connaissance sémantique de Logus.

Cette expérience a consisté à établir un lexique commun entre le corpus et la connaissance sémantique, puis d'analyser les positions verbales en fonction des types associés à chaque mot dans l'ontologie actuelle du système. L'analyse de corpus a mis en relief des phénomènes de variation sémantique qui, pour être pris en compte, nécessitaient l'aménagement de liens entre les types sémantiques. Les solutions proposées pour répondre à cette diversité sont la conversion de type et les propriétés saillantes, soit intervenir sur la structure lexicale afin de ne pas intervenir sur l'ontologie, ce qui permet de conserver l'arborescence de l'ontologie.

L'étude mériterait d'être approfondie et transposée à d'autres corpus afin de savoir dans quelle mesure le corpus de contes constitue un biais vis-à-vis des types d'interaction auxquelles sera confronté le robot à l'avenir. C'est dans cette perspective applicative que l'on pourra évaluer l'apport de cette recherche, c'est-à-dire à travers une comparaison d'un corpus d'interaction orale enfant-robot en situation réelle.

6 RÉFÉRENCES

Bassano D., Labrell F. et Champaud C. (2005). « Le DLPF, un nouvel outil pour l'evaluation du developpement du langage de production en francais ». Enfance, 2(5), p. 171-208.

Blanche-Benveniste C. (2005). « L'étude grammaticale des corpus de langue parlée en français ». Dans G. Williams (dir.), La linguistique de corpus. Rennes : PUR.

Devillers L., Bonneau-Maynard H., Rosset S., Paroubek P., Mostefa D., Choukri K., Charnay L., Bousquet C., Vigouroux N., Bechet F., Romary L., Antoine J-Y., Villaneau J., Vergnes M. et Goulian J. (2004). « The French Evalda-Media project: the evaluation of the understanding capabilities of Spoken Language Dialogue Systems ». Actes de LREC 2004. p. 2131-2134.

El Maarouf I. (2009). « Natural Ontologies at Work: investigating fairy tales ». Dans Actes CLC 2009.

C. Fabre et D. Bourigault (2006). « Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques ». Dans Actes de TALN'2006.

Firth J.R. (1957). « A synopsis of linguistic theory 1930-1955 ». Dans Studies in Linguistic Analysis. p. 1-32.

Hanks P. (2008). « Lexical Patterns: From Hornby to Hunston and Beyond ». Dans Actes d'Euralex 2008.

Hanks P. et Ježec E. (2008). « Shimmering lexical sets ». Dans Actes d'Euralex 2008.

Kilgarriff A., Rychly P., Smrz P. et Tugwell, D. (2004). « The Sketch Engine ». Dans Actes d'Euralex 2004. p. 105-116.

Lambert E. et Chesnet D. (2001). « Novlex: une base de données lexicales pour les élèves de primaire ». L'Année Psychologique 101, p. 277-288.

Lété B., Sprenger-Charolles L. et Colé P. (2004). « MANULEX : A grade-level lexical database from French elementary-school readers ». Behavior Research Methods, Instruments, & Computers, 36, p. 156-166.

Pustejovsky J. (1995). The Generative Lexicon. Cambridge : MIT Press.

Pustejovsky J., Hanks P. et Rumshisky A. (2004). « Automated Induction of Sense in Context ». Dans Coling 2004.

Pustejovsky J. et Ježek E. (2008). « Semantic Coercion in Language: Beyond Distributional Analysis ». Italian Journal of Linguistics, 20(2).

Saint-Aimé S., Le-Pévédic B., Duhaut D. et Shibata T. (2007). « Emotirob : Companion Robot Project ». Dans Actes de IEEE RO-MAN 2007, 16th IEEE Int. Symp. On Robot and Human Interactive Communication.

Schmidt H. (1994). « Probabilistic Part-of-Speech Tagging Using Decision Trees ». International Conference on New Methods Language Processing, p. 44-49.

Sinclair J. (1991). Corpus, concordance, collocation: Describing English language. Oxford : Oxford University Press.

Villaneau J., Antoine J-Y. et Ridoux O. (2002). « LOGUS : un système formel de compréhension du français parlé spontané - présentation et évaluation ». Dans Actes de TALN'2002. p. 165-174.

A CORPUS-BASED STUDY OF LATINATE WORDS IN CONTEMPORARY ENGLISH

Alex Chengyu Fang[1], Jing Cao[1] and Nancy Ide[1,2]
1. The Dialogue Systems Group, Department of Chinese, Translation and Linguistics, City University of Hong Kong, Hong Kong SAR, China
2. Vassar College, USA

ABSTRACT

The English language has borrowed extensively from Latin. Despite their widely acknowledged importance, however, Latinate words are under-studied especially quantitatively according to their use across different linguistic settings. In this paper, we report a corpus-based survey of Latinate words and their use in contemporary English. The objective is to chart the use and distribution of Latinate words across a set of different text categories and subject domains in order to identify patterns of variation across such settings. The British National Corpus is chosen for its large size and wide variety of texts. Our results will show that the density of Latinate words, calculated as the proportion of Latinate words amongst all running tokens, successfully separates speech from writing and, within writing, academic prose from non-academic prose. The correlation between Latinate density and degrees of text formality is then compared with human ranking and also analyzed by the linear regression analysis. The results suggest a strong correlation between Latinate density and degrees of text formality. Our results also show that even different domains have their own preferences for the use of Latinate words. Our investigation is significant in two respects. Firstly, it is probably the first corpus-based, large scale survey of the use of Latinate words in contemporary English. Secondly, it measures the density of Latinate words both across different text categories as an important stylistic feature and across a set of different domains as a subject-specific differentia. Such an investigation lends itself to our understanding of the impact of Latin on contemporary English and the empirical findings will also contribute to practical applications such as automatic text classification and genre detection.

1 INTRODUCTION

The English language has borrowed extensively from Latin. To this day, Latin words and words of a Latin origin still have an important presence in our daily spoken and written communication. The borrowing includes both scholarly and everyday words (e.g. Stockwell and Minkova, 2001) and many text types and subject domains are characterised by their extensive use of such words. For instance, Laar (1998) reports a high proportion of Latin component in English medical texts which contributes to the distinction between medical texts and texts of other kinds. De Forest and Johnson (2001) analyze the density of Latinate words in the speeches and letters of Jane Austen's characters. Their study shows that a higher density of Latinate words indicates a higher social status and education of the speaker and that a lower density indicates lesser intelligence or humble birth. A most recent study (Márquez, 2007) touches upon a similar observation in an attempt to define a core vocabulary by looking at the top one thousand most frequent words of British English. The study has nevertheless

shown that Latin is "not only a supplier of technical vocabulary" (Márquez, 2007: 712) but contributes in a significant way to the top one thousand of English. A much-quoted study in this area, i.e. Roberts (1965), reports the composite nature of borrowed words from languages such as Latin in American English but this study is based on a vocabulary list and does not refer to the actual use in natural texts. Despite their widely acknowledged importance, however, Latinate words are under-studied especially systematically according to pre-defined linguistic settings, such as types of writing or text categories, and quantitatively according to their authentic use in a large corpus of naturally occurring texts.

In this paper, we report a survey of Latinate words and their use in contemporary English. The objective of the survey, which is probably the largest and most extensive so far, is to chart the use and distribution of Latinate words across a set of different text categories and subject domains. We report their frequencies of use and present a quantitative description of their distribution in different text categories (such as writing vs. speech and academic prose vs. non-academic prose) and different domains (such as medicine and social sciences). The survey is significant in that it makes use of a large corpus of contemporary British English totaling one hundred million words. To our knowledge, no similar survey has ever been performed on such a scale. The survey is also significant in that it measures the use of Latinate words both across different text categories as an important stylistic feature and across a set of different domains as a subject-specific differentia. As our results will show, such a study not only lends itself to our understanding of the impact of Latin on contemporary English but will also contribute to practical applications such as automatic text classification and genre detection.

Our paper will be organised as follows. Section 2 will first discuss our methodology and introduce the resources (i.e. the corpus material and the lexical resource for etymological information), followed by the creation of the sub-corpora and the generation of a reference list of Latinate words. Section 3 will describe our experiment concerning the relationship between the chosen linguistic setting and the density of Latinate words defined as the proportion of Latinate word tokens amongst all the word tokens found for respective categories, while Section 4 will be devoted to the experiment on subject domains. We shall then draw some initial conclusions in Section 5.

2 METHODOLOGY AND RESOURCES

Our study attempts to extend the previous studies by examining quantitatively the density of the Latinate words across a wide range of naturally occurring texts. The density of Latinate words is used here to refer to the proportion of Latinate word tokens over the total number of word tokens in a given text category or domain. Since many text types and subject domains are characterised by their extensive use of Latinate words, we attempt to investigate the observable features of the density of Latinate words across different text categories and domains, and to find out whether such density can classify text categories in a meaningful way and whether different domains have their own preferences for the use of Latinate words. While our primary aim is to chart and characterise the use of Latinate words across the chosen categories and domains, we also hope that the empirical findings resulting from our investigation will also feed into practical natural language processing systems that perform automatic text categorisation and genre detection.

To achieve our objectives, two language resources are needed: 1) a large corpus of samples of contemporary English that is already encoded for text categories and subject domains, and 2) a large lexicon of contemporary English that provides etymological information for its component entries. For the former the British National Corpus (BNC) was selected for its

large size, representative categories and subject-specific domains and, for the latter, we used the machine readable Collins English Dictionary (CED) as the lexical resource for etymological information. The remainder of this section will introduce briefly the two resources.

2.1 Corpus Resource and Sampling

The British National Corpus (BNC), chosen as the basis of our study, is a large collection of samples totaling about 100 million word tokens designed to represent a wide cross-section of British English from the later part of the 20th century. It is encoded variously according to time, demography, medium, categories and domains. Table 1 summarizes the text classification of the BNC according to the *BNC User Reference Guide*.

Text Category		Token
Spoken	Conversation	4,233,962
	Other Speech	6,175,896
Written	Academic Prose	15,781,859
	Fiction	16,143,913
	Newspapers	9,412,174
	Non-academic Prose	24,179,010
	Other Published Writing	17,970,212
	Unpublished Writing	4,466,681
Total		98,363,707

Table 1: Text classification of the BNC

As shown in Table 1, the BNC contains both a spoken section (about 10% of the total corpus size) and a written section (about 90% of the total corpus size), spread over eight different text categories, which represents an ideal setting for our intended task of examining features of Latinate words across text categories.

2.1.1 THE CREATION OF A SUB-CORPUS FROM THE BNC FOR TEXT CATEGORIES

A sub-corpus was created with samples randomly selected from the eight text categories in the BNC. A total of 3 million word tokens were selected for each category at the text level, and Table 2 presents the actual tokens sampled for each category.

Text Category		Text Code	Token
Spoken	Conversation	CONV	3,017,930
	Other Speech	OSP	3,019,043
Written	Academic Prose	AC	3,124,550
	Fiction	FIC	3,026,196
	Newspapers	NEWS	3,018,301
	Non-academic Prose	NONAC	3,083,486
	Other Published Writing	OPUB	3,013,586
	Unpublished Writing	UNPUB	3,001,746
Total			24,304,838

Table 2: Tokens sampled in text categories

Of this sub-corpus, 80% is used as the training set, and the remaining 20% as the test set. Our primary data comes from the training set. In the event of any significant findings, we use the test set to verify them. Table 3 shows the basic statistics of our training, and they are represented in descending order according to the type-token ratio (TTR).

Text Code	Token	Type	TTR
NONAC	2,451,482	71,460	2.91
OPUB	2,354,825	66,269	2.81
NEWS	2,360,843	64,338	2.73
AC	2,468,802	61,990	2.51
UNPUB	2,395,601	53,404	2.23
FIC	2,382,786	40,966	1.72
OSP	2,382,061	26,590	1.12
CONV	2,368,324	19,775	0.83

Table 3: Basic stats of the training set

Two features emerge from Table 3. First, there seems to be a clear cut between the spoken and written texts, when OSP and CONV are grouped together at the bottom of the scale. Second, all the written texts have a similar type-token ratio except FIC which shows a closeness to spoken texts. It is well understood in corpus-based linguistic studies that the written genre tends to have higher vocabulary content, indicated by the type-token ratio, than the spoken genre.

2.1.2 THE CREATION OF SUB-CORPUS FROM THE BNC FOR SUBJECT DOMAINS

The BNC has eight text categories (Table 1), from which the category of academic prose (AC) was chosen as the basis of our experiment to investigate the use of Latinate words in relation to subject domains. AC comprises six subject domains, for each of which 500,000 word tokens were randomly sampled to form the sub-corpus. See Table 4.

Subject Domain	Domain Code	Token
Humanities and arts	HUM	524,224
Medicine	MED	504,857
Natural science	NAT	536,499
Politics, education and law	POL	511,935
Social science	SOC	511,655
Technology, computing and engineering	TEC	535,380
Total		3,124,550

Table 4: Tokens sampled in subject domains

Again, training and test sets were divided with a ratio of 8 to 2. Table 5 shows the basic statistics of the training set, and they are represented in descending order according to the type-token ratio (TTR).

Domain Code	Token	Type	TTR
NAT	429,120	21,096	4.92
HUM	410,281	21,226	5.17
MED	402,025	18,966	4.72
SOC	401,346	16,532	4.12
POL	407,480	14,124	3.47
TEC	418,550	13,196	3.15

Table 5: Basic stats of the training set

2.2 Lexical Resource and Reference List

The Collins English Dictionary (CED is chosen for three main reasons. First of all, the CED is a well-recognised dictionary with etymological information in addition to information such as part of speech, pronunciation, and sense definition. Secondly, the LDC offers the CED in a machine readable mode (ACL/DCI, 1993). In other words, the LDC version of CED is not just a machine readable dictionary but is also parsed into different fields of information for further applications such as information retrieval and natural language processing. Finally, the CED is primarily a British English dictionary and thus well suited to our work on the British National Corpus, which is also British English.

From the CED, a reference list was generated that contains all the headword entries with a Latinate origin. When generating the reference list, we first counted entries with etymological information, both explicit and implicit. The CED has 249,331 entries, among which 28,526 contain explicit etymological information, i.e. entries marked with `ety`. Implicit etymological information was recovered for an additional 20,066 entries via a label (`head`) that links the derived entries to their corresponding root with an explicit etymology indication `ety`. For example, the entry *blandness* does not have an explicit etymology indication. However, this entry has a `head` field linking the entry to *bland*, which has an explicit indication of its Latin origin. The entry *blandness* is thus counted as a Latinate word. Therefore, the total number of entries with etymological information, both explicit and implicit, is up to 48,593, accounting for 19.5% of the CED entries. For all the 48,593 `ety`-entries, 247 different language origins are coded. See Table 6 for the top 20 language origins together with the raw counts and associated proportions in the `ety`-entries.

Origin	Freq.	%
Latin	11,711	24.10
Old English	7,784	16.02
Old French	7,439	15.31
French	3,904	8.03
New Latin	3,102	6.38
Greek	2,109	4.34
Late Latin	1,797	3.70
Medieval Latin	1,292	2.66
Italian	983	2.02
German	643	1.32
Spanish	636	1.31
Old Norse	634	1.30
Scandinavian	480	0.99
Middle Dutch	369	0.76
Dutch	349	0.72
Anglo-French	347	0.71
Hindi	325	0.67
Arabic	244	0.50
Hebrew	205	0.42
Middle Low German	193	0.40

Table 6: Top 20 etymological origins in the CED

It is also noticeable in Table 6 that among the top 20 languages, there are four sub-divisions of Latin origin, namely, Latin, New Latin, Late Latin and Medieval Latin. Actually, among the 247 different languages coded in the CED, there are nine sub-divisions of Latin origin and Table 7 presents the types of sub-divisions, the raw counts and their associated proportions in the `ety`-entries. Given the objectives of the current study, we conflated all the

sub-divisions into 'Latin' and thus all entries originating from Latin come up to 17,943, which were subsequently included in the reference list to be used for the identification of the Latinate words in the BNC.

Sub-divisions of Latin Origin	Freq.	%
Latin	11,711	24.100
New Latin	3,102	6.384
Late Latin	1,797	3.698
Medieval Latin	1,292	2.659
Anglo-Latin	32	0.066
Old Latin	3	0.006
Ecclesiastical Latin	2	0.004
Modern Latin	2	0.004
Vulgar Latin	2	0.004
Total	17,943	36.93

Table 7: Latin entries in the CED

A stoplist of 2,000 most frequent words was created from the BNC and any item covered by the list was excluded from the reference list and therefore from the frequency counts reported in this article. We believe that the 2,000 most frequent words can be regarded as part of the core vocabulary and that the words of Latin origin in the stoplist tend to have the feature of 'nativeness' and therefore can be regarded as native words.

3 INVESTIGATING LATINATE WORDS IN THE BNC ACCORDING TO CATEGORIES

Our corpus-based investigation into the density of Latinate words is made in two parameters: by text categories and by subject domains. This section is devoted to the investigation of Latinate density across different text categories, and section 5 will discuss the Latinate density across different domains.

3.1 Basic Statistics of Latinate Words in Text Categories

From the training set of the subcorpus, we extracted all the lemmatised word tokens (or the *headwords* in the BNC) that were matched with the reference list of Latinate words generated from the CED. Such a headword list was then filtered with the top 2,000 stoplist. Table 8 presents the final statistics of Latinate words in terms of type, token and type-token ratio. As is shown, the text categories are arranged by the type-token ratios (TTRs) of the Latinate words in descending order.

Text Code	Latinate Token	Latinate Type	TTR
CONV	11,343	1,455	12.83
FIC	39,863	3,446	8.64
OSP	28,825	2,487	8.63
NEWS	45,198	3,033	6.71
OPUB	55,153	3,642	6.60
UNPUB	52,850	2,965	5.61
NONAC	80,823	4,145	5.13
AC	103,112	4,190	4.06

Table 8: Basic Latinate stats

As noted above, the data shows that the spoken texts tend to have a higher TTR of Latinate words with CONV at the top of the list followed by fiction, which is known to have a greater assimity to speech than to writing. The written texts are more likely to have a lower TTR of Latinate words; for example, AC is at the bottom of the list of TTR.

3.2 Ranking of Categories by Latinate Density

We then computed the Latinate density, *D*, by calculating the proportion of tokens of Latinate words over the total tokens in each text category:

$$D = \frac{\text{number of Latinate word tokens}}{\text{number of total word tokens}} \times 100 \qquad (1)$$

Table 7 presents the eight BNC text categories sorted according to *D* in ascending order, and such a ranking is also defined as R_{d1}, as shown in Column 1.

R_{d1}	Text Code	Word Tokens	Latinate Tokens	*D*
1	CONV	236,8324	11,343	0.48
2	OSP	238,2061	28,825	1.21
3	FIC	2,382,786	39,863	1.67
4	NEWS	2,360,843	45,198	1.91
5	UNPUB	2,395,601	52,850	2.21
6	OPUB	2,354,825	55,153	2.34
7	NONAC	2,451,482	80,823	3.30
8	AC	2,468,802	103,112	4.18

Table 7: Ranking of categories by Latinate density

As noted above, AC has the highest density of Latinate words, 4.18%, whereas CONV has the lowest density of 0.48%. From the viewpoint of speech and writing, we notice that the written texts are grouped together towards the bottom of the scale and that the spoken texts are clustered together at the top of the scale. In other words, writing in general exhibits a higher Latinate density than speech. Moreover, within the six written categories, AC has a higher density than NONAC. Published writing, such as AC, NONAC and OPUB, has a higher density than unpublished writing, UNPUB. Fiction (FIC) has the lowest density among the written texts, boarding the spoken texts on the scale represented in Figure 1, which corresponds neatly with indications in terms of type-token ratio described in Section 3.1, suggesting a closer affinity with the spoken genre.

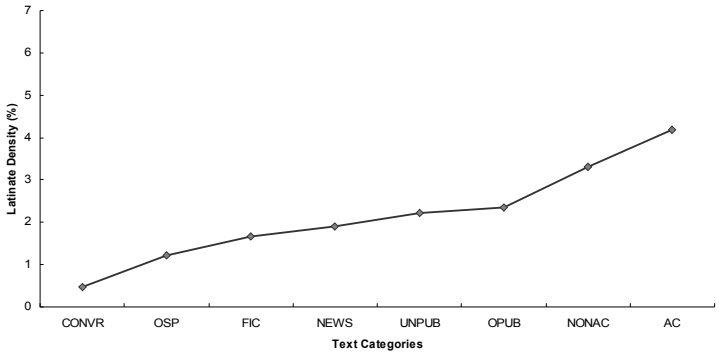


Figure 1: Latinate Density of the Training Set

3.3 Comparison with Human Ranking of Categories

3.3.1 HUMAN RANKING

Seven human subjects (six PhD students and one professor in linguistics) were invited to evaluate the formality of the eight text categories independently. They were asked to rank the text categories in the order of formality by specifying 1, 2, 3…etc, with 1 being the most informal and 8 the most formal. Inter-rater reliability was then tested by computing the intra-class correlation (ICC) coefficient. The value of the ICC coefficient is 0.857 with $p<0.001$, which is considered as outstanding inter-rater reliability (Landis and Koch, 1977). Next, the means of the human judgments were computed, according to which the eight different text categories were ranked. See Table 8 for the results with R_{h1} indicating manual ranking.

3.3.2 RANK COMPARISON

The ranking according to Latinate density (R_d) was then examined by comparing it with the human ranking (R_{h1}). The absolute difference of each paired rankings (*Difference*) was calculated and Table 8 presents the results.

Text Code	R_{d1}	R_{h1}	*Difference*
CONV	1	1	0
OSP	2	2	0
FIC	3	4	1
NEWS	4	6	2
UNPUB	5	3	2
OPUB	6	7	1
NONAC	7	5	2
AC	8	8	0

Table 8: Density vs. human rankings.

Based on *Difference*, the Spearman rank correlation coefficient r_s was calculated between the ranking of Latinate density and human ranking according to the formula:

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2 - 1)} \qquad (2)$$

As a result, the value of the Spearman rank correlation coefficient is 0.833, which is significant at the level of 0.02. In other words, there is strong evidence of agreement between the ranking of Latinate density and the human ranking. This result also suggests a strong correlation between Latinate density and text formality.

3.4 Linear Regression Analysis

Motivated by the substantial agreement between the density and human rankings, a linear regression analysis was further performed to examine the relation between Latinate density and text categories. Figure 2 is a graphical representation of the analysis, where the *Y*-axis represents Latinate density and the *X*-axis the eight text categories arranged in the same manner as in Table 7. As can be seen, the points seem to follow a linear pattern with a positive slop. The linear correlation coefficient *r* (=0.971) suggests a strong positive linear relationship between Latinate density and degree of text formality, and the coefficient of determination (r^2) is 0.942, indicating that about 94.2% of the variation in the density data can be explained by the degree of text formality.

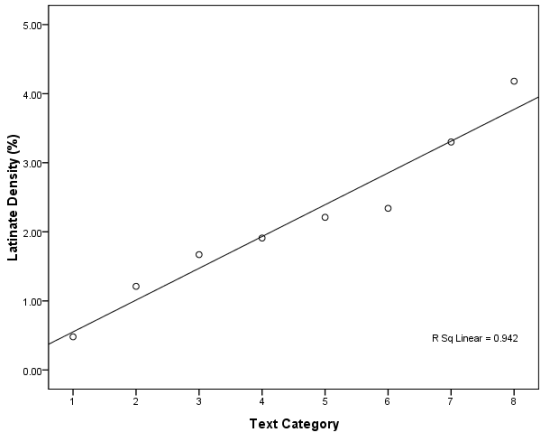


Figure 2: A linear regression analysis

To sum up, our data shows that Latinate density seems to be a stylistic characteristic correlating with degrees of formality, which distinguishes not only speech from writing as two broad genres but also informal categories from the formal ones within the written genre. Similar empirical results were obtained for the test set.

4 INVESTIGATING LATINATE WORDS IN THE BNC ACCORDING TO DOMAINS

4.1 Basic Statistics of Latinate Words in Domains

Table 9 summarises the basic statistic data and the domains are presented in descending order according to the type-token ratio (TTR) of their Latinate words. One obvious feature emerging from Table 10 is that domains are separated into two groups: arts (HUM, POL and SOC) and sciences (NAT, MED, and TEC), quite unlike our previous observation based on word token TTR in Section 2.1.2. Domains belonging to arts have a higher Latinate TTR, all above 12%, while domains belonging to sciences have a comparatively lower TTR, all below 9%. This phenomenon can be explained by the observable fact that the sciences domains seem to make a heavier use of Latinate words than the arts domains, thus yielding some initial indication of a different degree of preference for Latinate words between arts and sciences.

Domain Code	Latinate Token	Latinate Type	Latinate TTR
HUM	14,750	2,159	14.64
POL	10,803	1,523	14.10
SOC	11,350	1,431	12.61
NAT	22,619	2,018	8.92
MED	24,915	1,846	7.41
TEC	18,675	1,116	5.98

Table 9: Basic Latinate stats for domains

4.2 Ranking of Domains by Latinate Words

Table 10 summarises the Latinate densities (*D*) for the six subject domains in the training set arranged according to *D* in ascending order.

R_{d2}	Domain Code	Total Tokens	Latinate Tokens	*D*
1	POL	407,480	10,803	2.65
2	SOC	401,346	11,350	2.83
3	HUM	410,281	14,750	3.60
4	TEC	418,550	18,675	4.46
5	NAT	429,120	22,619	5.27
6	MED	402,025	24,915	6.20

Table 10: Ranking of domains by Latinate density

As can be seen from Table 10, MED has the highest Latinate density of 6.20% and POL has the lowest density of only 2.65%. From the view point of arts and sciences, science domains (namely, MED, NAT and TEC) have a comparatively higher density than arts (namely, HUM, SOC and POL). Within the science domains, MED has a higher proportion of Latinate words than TEC; among the arts, HUM has a higher proportion of Latinate words than POL. Figure 3 illustrates the correlation between domains and Latinate density.

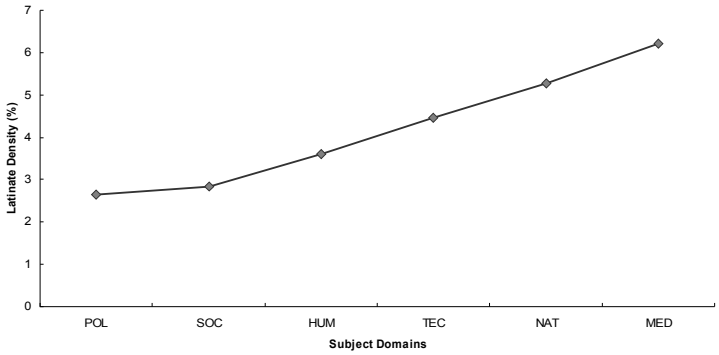


Figure 3: Latinate density of the training set

4.3 Comparison with Human Ranking of Categories

4.3.1 HUMAN RANKING

The same seven human subjects were also asked to rank the subject domains in the order of formality by specifying 1, 2, 3…etc, with 1 being the most informal and 6 the most formal. Inter-rater reliability was then tested by computing the intra-class correlation (ICC) coefficient. The value of the ICC coefficient is 0.616 with $p<0.001$, which is considered as

acceptable (Landis and Koch, 1977). Again, the means of the human judgments were computed, according to which the six subject domains were ranked (See R_{h2} in Table 11).

4.3.2 RANK COMPARISON

The ranking according to Latinate density (R_{d2}) was then examined by comparing it with the human ranking (R_{h2}). The absolute difference of each paired rankings (*Difference*) was calculated and Table 11 presents the results.

Text Code	R_{d2}	R_{h2}	*Difference*
POL	1	3	2
SOC	2	2	0
HUM	3	1	2
TEC	4	4	0
NAT	5	6	1
MED	6	5	1

Table 11: Density vs. human rankings

Based on *Difference*, the Spearman rank correlation coefficient r_s was calculated between the ranking of Latinate density and human ranking in the setting of subject domains. The value is 0.714, which is not significant. In other words, there is very weak evidence of agreement between the two rankings. Yet, it is still noticeable that both rankings have classified the same domains into two broader categories: Arts and Sciences.

Due to the insignificant correlation, no further linear regression test was made. Nevertheless, the empirical results still show that even different domains have their own preferences for the use of Latinate words. Subject domains of sciences tend to have a higher proportion of Latinate words than those of arts. In this sense, it is safe to say that the use of Latinate words is correlated to subject domains. Similar results were also found in the test set.

5 CONCLUSION

We described an investigation of Latinate words which is significant in two respects. Firstly, it is the most extensive study of Latinate words based on a large corpus of contemporary British English. Secondly, we investigated the use of Latinate words in a linguistic setting that involved not only a spectrum of text categories ranging from informal speech to formal academic writing but a variety of subject domains in arts and sciences. The machine-readable Collins English Dictionary was used as our lexical resource for a reference list of Latinate words. The British National Corpus was used as the basis of the study for authentic texts. A sub-corpus was created with a total of three million words for each of the eight text categories. A second sub-corpus of six subject domains was created from academic prose with a total of 500,000 word tokens for each domain, totaling three million word tokens.

Latinate density was calculated as the total number of Latinate word tokens over the total number of word tokens, first across the eight different text categories and then across the six subject domains concerned in the reported study. The results from primary data sets were evaluated against the results from the secondary data sets.

Our findings show that there is an uneven use of Latinate words across the text categories. To be more exact, Latinate density can be used to distinguish speech from writing, and moreover, between formal and informal writing. The density of Latinate words therefore suggests that this measure can be used as a stylistic characteristic that relates unambiguously

to degrees of formality with good potentials for application in natural language processing systems to classify texts and to detect novel genres.

Our investigation also shows that even different subject domains have their own preferences for the use of Latinate words. Domains in the sciences have a higher proportion of Latinate words than those in arts. The findings indicate that the density of Latinate words can be possibly used as a subject-specific differentia, for the separation of texts in arts and sciences at least.

Our survey thus demonstrates on an empirical basis that the use of Latinate words not only distinguishes texts on a scale of different formalities but that different domains seem to have a different proportion and therefore preference for the use of Latinate words, a finding that will contribute to applications in automatic text classification and genre detection, a promising potential that we are currently investigating in a separate study.

6 ACKNOWLEDGEMENTS

The research reported in this article was support in part by research grants (7002190, 7200120 and 7002387) from City University of Hong Kong. The authors would like to acknowledge valuable input and assistance received from the members of the Dialogue Systems Group at the Department of Chinese, Translation and Linguistics, in particular, Claire Li, John Hanhong Li, and Maggie Xing Zhang.

7 REFERENCES

ACL/DCI. (1993). *Linguistic Data Consortium.* Philadelphia.

Roberts A. H. (1965). *A Statistical Linguistic Analysis of American English.* The Hague: Mouton.

BNC User Reference Guide. URL: http://www.natcorp.ox.ac.uk/XMLedition/URG/index.html

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.

Lee D. (2001). « Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. » *Language Learning & Technology,* 5(3), p. 37-72.

Laar M. (1998). *The Latin Component in English Medical Texts and Some of the Possibilities It Offers for Interdisciplinary Integrated Teaching.* Jyväskylä.s. p. 171-174.

Forest D. M. and Johnson E. (2001). « The Density of Latinate Words in the Speeches of Jane Austen's Characters. » *Literary and Linguistic Computing*, 16(4), p. 389-401.

Márquez M. F. (2007). « Renewal of Core English Vocabulary: A Study Based on the BNC. » *English Studies*, 88(6), p. 699-723.

Peters P. (1998). « Surveying Contemporary English Usage. » *English Today* 56, 14 (4), p. 3-6.

Stockwell R. and Minkova D. (2001). *English Words: History and Structure.* Cambridge: Cambridge University Press.

DISCOURS ÉVALUATIF : UNE CAMPAGNE D'ANNOTATION POUR LA VALIDATION DE PATRONS

Stéphane Ferrari[1], Thierry Charnois[1], Agata Jackiewicz[2], Pierre Gardin[1] et Antoine Widlöcher[1]
1. GREYC UMR 6072 – Université de Caen - Basse Normandie
2. LALIC – Université Paris-Sorbonne

RÉSUMÉ

Dans cet article, nous nous intéressons au discours évaluatif, plus particulièrement aux énoncés exprimant des jugements d'évaluation. Nous présentons une première campagne d'annotation visant à permettre l'évaluation d'outils automatiques d'analyse de ces phénomènes. Elle s'appuie sur une étude linguistique menée au préalable sur un corpus extrait du journal Les Échos, constitué essentiellement de portraits de personnages de la vie économique, qui a permis de dégager des régularités dans l'expression des jugements d'évaluation portant sur des personnes au sein de constituants détachés. Nous présentons brièvement la méthode informatique mise en œuvre pour permettre de repérer automatiquement ces régularités et les résultats obtenus sur un deuxième corpus, constitué de portraits issus cette fois du journal Le Monde. Malgré quelques différences de genre, les mêmes régularités paraissent en grande partie efficaces pour détecter des opinions. Nous proposons en conséquence une première campagne d'annotation manuelle pour l'évaluation quantitative de ce type d'outil. Nous présentons le protocole et le logiciel utilisés pour l'annotation ainsi que les taux d'accord entre annotateurs à l'issue de deux expérimentations successives. Les résultats obtenus montrent des difficultés à produire des annotations complexes fiables, mais une possibilité cependant d'en produire de simples avec des taux d'accord élevés.

1 INTRODUCTION

Le discours évaluatif connaît un intérêt croissant depuis une dizaine d'années, notamment en Traitement Automatique des Langues (TAL), pour des applications possibles dans des domaines telles que la veille d'opinion, l'analyse de tendances ou de marchés. Les travaux relatifs menés en informatique traitent essentiellement de la fouille d'opinions et de l'analyse de sentiments. Ils peuvent être classés selon trois axes principaux : (i) constitution de ressources lexicales pour la fouille d'opinion ; (ii) classification de textes d'opinions ; (iii) analyse d'opinion dans les textes. Proche de cette dernière problématique, notre approche vise, à terme, la mise en place d'outils automatiques d'analyse sémantique fine des opinions et des jugements exprimés au sein de documents textuels numériques. Plus précisément, nous suivons un point de vue multidisciplinaire, consistant à croiser les regards de linguistes et d'informaticiens pour proposer des modèles opératoires qui s'appuient en amont sur une expertise de la langue. En nous inspirant notamment du modèle de l'Appraisal (Martin et White, 2005), nous visons l'analyse sémantique d'énoncés évaluatifs en français pour rendre compte des différentes propriétés relatives aux opinions exprimées. L'étude présentée dans

cette communication concerne plus particulièrement l'analyse de certaines formes récurrentes d'énoncés liés à l'expression de jugements d'évaluation portant sur des personnes, l'automatisation de cette analyse par une chaîne de traitement et la problématique de son évaluation sur corpus.

Nous présentons le cadre de notre étude et un bref état de l'art des travaux connexes menés en informatique. Nous exposons ensuite la méthode qui a été suivie pour repérer des patrons linguistiques dans un premier corpus, en proposer une mise en œuvre informatique et tester leur pertinence sur un deuxième corpus. Enfin, au vu des résultats obtenus, nous détaillons les premières phases de la campagne d'annotation que nous menons en vue de la mise en place de ressources de référence pour l'évaluation : le protocole proposé, l'outil utilisé pour l'annotation manuelle, puis les taux d'accords mesurés entre annotateurs à l'issu de deux expérimentations successives. Nous concluons en précisant quelles perspectives sont désormais envisagées.

2 CADRE, TRAVAUX CONNEXES

En TAL, notre étude s'inscrit dans le cadre de l'analyse automatique de l'expression de l'évaluation dans le discours. Ce thème est en émergence depuis une dizaine d'années, en raison de ses nombreuses applications potentielles : veille industrielle, analyse de tendances, sondages d'opinion pour la consommation, la politique… Une monographie assez complète sur le sujet a été réalisée par Pang et Lee (2008), précisant les différentes tâches que l'analyse d'opinion cherche à résoudre, les méthodes développées afin d'atteindre ces objectifs, ainsi que les applications existantes et potentielles de ces techniques. Les différentes tâches que visent la plupart des applications informatiques s'inscrivent dans trois catégories principales : (i) la constitution de ressources, essentiellement des lexiques pour la fouille d'opinion ; (ii) la classification de documents ou de parties de documents, essentiellement pour en caractériser la polarité ou simplement le caractère subjectif ; (iii) l'analyse locale de l'expression d'opinion dans les textes.

12. Les approches visant la constitution de ressources dépendent des propriétés lexicales étudiées : le caractère subjectif ou objectif (Baroni et Vegnaduzzo, 2004 ; Riloff et Wiebe, 2003), la polarité positive ou négative (Hatzivassiloglou et McKeown, 1997 ; Turney et Littman, 2003), ou encore une caractérisation plus fine du sens de termes comme dans (Wiebe et Mihalcea, 2006) et dans les travaux relatifs à SentiWordNet (Esuli et Sebastiani, 2006). Quelques travaux s'inspirent de la théorie de l'Appraisal (Martin et White, 2005) pour caractériser des expressions complexes, comme dans (Whitelaw *et al.*, 2005).
13. De nombreux travaux visent la catégorisation de textes ou parties de textes, en termes de polarité globale (texte positif, négatif ou neutre), parfois simplement de subjectivité. La campagne d'évaluation DEFT07 (Grouin *et al.*, 2009) avait cet objectif, pour la langue française. Différents domaines applicatifs sont concernés : critiques de films (Turney, 2002 ; Pang et Lee, 2004) et une partie de DEFT07, textes politiques dans une autre partie de DEFT07, critiques de produits (Hu et Liu, 2004), *etc*. Les techniques utilisées sont variées, majoritairement issues des domaines de la fouille de données et de l'apprentissage automatique.
14. Une série de travaux visent des analyses plus locales, pour distinguer les opinions des faits dans un système de Question/Réponse (Yu et Hatzivassiloglou, 2003), pour proposer un résumé des points sur lesquels portent les critiques émises par les consommateurs dans les travaux de (Hu et Liu, 2004), ou encore pour annoter en contexte les expressions d'opinion. Les annotations dépendent alors des applications visées, incluant la détermination du caractère subjectif ou de la

polarité, mais aussi d'autres caractéristiques comme les cibles (objets des évaluations), les sources (émetteurs), l'intensité, *etc*. (Wiebe *et al.*, 2005 ; Read *et al.*, 2007 ; Ferrari *et al.*, 2009).

Notre propre approche s'inscrit dans le cadre de cette dernière catégorie, dans la mesure où nous visons à terme une analyse locale d'expressions d'opinion. Pour cela, nous nous appuyons sur une étude linguistique d'un phénomène spécifique, présentée dans la suite.

3 PATRONS LINGUISTIQUES

Une étude préliminaire a porté sur le phénomène discursif du détachement et, plus spécifiquement, l'expression d'opinion au sein de constituants périphériques, extraprédicatifs (Jackiewicz, 2009). Elle a permis de mettre en évidence une vingtaine de patrons linguistiques, et a donné lieu à une première expérimentation informatique dans la plate-forme Linguastream (Widlöcher et Bilhaut, 2008) pour en permettre l'évaluation sur des extraits du journal Le Monde. Avant de décrire la mise en œuvre informatique des patrons, nous présentons brièvement l'étude linguistique menée en corpus sur les constituants détachés. Pour plus de détails sur les aspects linguistiques, nous renvoyons le lecteur à (Jackiewicz, 2009) et (Jackiewicz *et al.*, 2009).

3.1 Constituants détachés et patrons linguistiques : présentation

Les éléments périphériques à la prédication principale dénotent souvent des appréciations, sous formes réduites et clairement délimitées, telles qu'illustrées dans les exemples 1 à 4 :

4. *Mais, **en politicien expérimenté**, élu pour la première fois à la Knesset il y a trente-cinq ans, il a su résister aux roquettes de ses adversaires politiques.*

1. ***Ni trop sentimental, ni trop énergique,** il maîtrise, **avec une finesse quasi mozartienne,** un lyrisme généreux.*

    - ***Militant mais opportuniste, franc-tireur mais habile, sociable mais anticonformiste,** le directeur de l'Opéra de Paris sait manier les paradoxes pour parvenir à ses fins.*

    - ***Figure légendaire de l'opposition au régime communiste, éminent professeur d'histoire médiévale,** ministre des affaires étrangères de la Pologne de 1997 à 2000, Bronislaw Geremek avait été élu au Parlement européen en 2004.*

Notre travail s'appuie sur des recherches sur l'apposition, les constructions détachées et les compléments circonstanciels décrites dans (Combettes, 1998 ; Neveu, 2000 ; Guimier, 1996). Au plan sémantique et énonciatif, nous empruntons l'appareil conceptuel de la théorie de l'Appraisal (Martin et White, 2005). Du point de vue discursif, nous analysons les relations rhétoriques qui peuvent s'ancrer sur un jugement évaluatif.

Pour reconnaître un constituant périphérique potentiellement axiologique, un ensemble de critères peut être mis en avant, dont les principaux sont les suivants (cf exemple 5) : délimitation par des signes de ponctuation, possibilité d'effacement, (relative) liberté de position, statut de prédication seconde, existence d'un référent sous-jacent…

- **Intuitif**, Joel Séché pressent l'essor du développement durable et anticipe le durcissement de la réglementation.

L'étude de ces constituants, menée manuellement sur des textes du journal Les Échos, nous a conduit à constituer manuellement une vingtaine de patrons linguistiques (ainsi qu'un lexique d'environ 550 termes) correspondant à 3 principales catégories :

- Constructions détachées :

2. Groupes adjectivaux (*imprévisible et fantasque, X*),
3. Constructions absolues (*l'œil vigilant*, X) ,
4. Participes (*réputé pour son caractère bourru, X*) .

- Compléments circonstanciels :
  6. Adverbes (*courageusement, X*),
  7. Groupes prépositionnels (*en mauvaise posture, X*) .
- Groupes nominaux avec ou sans déterminants :
  9. N de GN... (*Femme de tête, X ; X, le maestro de la désinflation*).

3.2 Mise en œuvre et test sur un second corpus

La poursuite de cette première étude a consisté en une extension et une évaluation sur de nouvelles données. Pour mener à bien ce travail, nous avons constitué un nouveau corpus, d'un genre proche de celui ayant permis l'analyse linguistique, regroupant 884 articles de type « Portrait » ou « Biographie » issus du journal Le Monde de la période juillet à décembre 2002. Le caractère plus généraliste de cette publication fait apparaître des différences dans le genre considéré, les personnes sur lesquelles portent les opinions pouvant être autant issues du monde de la politique ou de l'économie que du milieu des arts et de la culture. Des évaluations y sont bien présentes sous les formes que nous avons étudiées, mais elles relèvent désormais davantage de l'appréciation sur des objets et plus seulement sur des jugements à l'égard de personnes ou de leurs actes.

La mise en œuvre de 10 patrons représentatifs parmi les 20 observés a été intégrée sous forme de règles Prolog dans un composant d'une chaîne de traitements Linguastream (Bilhaut et Widlöcher, 2006 ; Widlöcher et Bilhaut, 2008), chaîne incluant une série de pré- et post-traitements (segmentation, catégorisation grammaticale à l'aide du Treetagger (Schmid, 1994), puis transformation des annotations au format HTML pour permettre une visualisation aisée avec un navigateur Web). Leur application sur ce deuxième corpus mène aux résultats suivants. 1966 patrons ont été repérés dans le corpus, dans 580 des 884 articles (65%). Leur répartition montre que 395 articles contiennent plus d'un patron, 107 plus de 5, 21 seulement plus de 10 ; le maximum de patrons par article est 23. La distribution sur le corpus est donc assez hétérogène. Nous renvoyons à (Jackiewicz *et al.*, 2009) pour une analyse plus détaillée de ces résultats, évoqués ci-après.

L'observation qualitative des résultats permet de dégager plusieurs pistes intéressantes pour notre étude. Nous constatons une forte concentration de patrons au sein de quelques articles, quand 35% n'en contiennent aucun, ce que nous pouvons partiellement expliquer par des préférences concernant les modes d'expression qui seraient propres aux auteurs. Les patrons concernant les groupes adjectivaux sont, sans surprise, les plus fréquemment rencontrés. Ceci va dans le sens des nombreux travaux de la communauté TAL qui concentrent leurs efforts sur les adjectifs. D'autres phénomènes intéressants apparaissent : présence de multiples constituants détachés successifs, des phrases averbales (7-8), des connecteurs introduisant des jeux d'opposition ou de renforcement (9), des modalités temporelles (*et, tantôt, sans, à la fois, mais, toujours, jamais,* etc.).

1. L'acteur d'Anouilh et de Chabrol, de Diderot et de Pinter, reste un voyageur traqué de l'âme. Inquiétant, ambigu, fascinant. (…). Laconique et inquiétant, discret et courtois, il cultive l'anonymat jusqu'à l'étrangeté.

- Froidement, avec méthode, en chemise blanche et cravate noire.

Réputée rigide, parfois cassante, elle est toujours difficile à manœuvrer. (..) Décriée, souvent méprisée par les caciques du parti, elle jouit malgré tout d'une véritable popularité parmi les militants.

Nous avons pu aussi constater des situations d'échec, des erreurs qui sont soit du bruit (des faux positifs), soit du silence (des oublis). En particulier, les cibles des opinions repérées peuvent être d'une nature autre que celle attendue (10). Ceci tient à la nature du journal dont est issu ce corpus, qui recèle désormais, même au sein de la catégorie « Portraits », des appréciations pouvant porter sur des objets, et non plus uniquement des jugements de personnes ou de comportements. C'est notamment le cas des personnes du monde des arts et de la culture, dont la production est jugée autant sinon plus qu'eux-mêmes en tant qu'individus.

1. Pour ses premiers romans, il pouvait " passer la nuit sur une phrase " prise, reprise, démantibulée, reconstruite, abandonnée, restructurée. "

L'énumération de patrons est reconnue, mais la cible n'est pas une personne.

Ce type d'erreur reste sans gravité au vu de notre objectif à long terme : nous repérons une catégorie différente d'opinions, dont le typage ne peut visiblement être effectué qu'en analysant précisément les cibles, perspective que nous envisageons à moyen terme. En revanche, certaines erreurs sont définitivement inacceptables :

- Difficile, pourtant, de l'ignorer : …

La forme de surface correspond à un patron, mais il n'y a ici aucun constituant périphérique.

- Bel homme, cultivé, surnommé " Harrison Ford " par ses camarades femmes qui apprécient son charme, Guglielmo Epifani arrive avec son staff ...

Exemple de silences multiples : seul *cultivé* est repéré par notre chaîne, à la fois à cause d'erreurs issues des prétraitements et de patrons trop restreints.

Nous avons donc décidé de mettre en place une campagne d'annotation manuelle afin de permettre une évaluation quantitative de tous ces phénomènes.

4 CAMPAGNE D'ANNOTATIONS

L'annotation que nous envisageons a pour finalité de permettre le calcul de taux de précision et de rappel pour les analyses automatiques développées. Nous présentons ici le début de la campagne d'annotation manuelle réalisée à cet effet. Elle se déroule sur une partie du corpus extrait du journal Le Monde. Nous décrivons ci-après le protocole que nous avons retenu pour deux phases successives, ainsi que l'outil utilisé : la plate-forme Glozz. Nous commentons les taux d'accord obtenus entre annotateurs pour ces deux phases et en tirons des conséquences pour la poursuite de cette campagne.

4.1 Protocole et outil d'annotation

4.1.1 PROTOCOLE PROPOSÉ AUX ANNOTATEURS

Le protocole d'annotation proposé pour lancer la campagne est, en substance, le suivant. Il a été agrémenté d'instructions relatives à l'usage de la plate-forme Glozz (voir infra), ainsi que d'une série d'exemples (Tableau 1) et de critères regroupés sous forme de tableaux pour préciser la nature des constituants repérés parmi les patrons précédemment exposés. Nous en rappelons l'essentiel ici.

Il était demandé aux annotateurs d'annoter des constituants périphériques (CP dans la suite) de deux natures : selon que le CP porte une valeur axiologique ou semble uniquement qualifier la cible sans porter d'opinion, choisir le type d'unité à annoter, **cp_axio** ou **cp_qua**. La délimitation du CP devait se faire en plaçant le début d'unité juste avant le premier mot et la fin d'unité juste après la ponctuation qui suit le CP (une virgule, un point final, une double ponctuation, des points de suspension). L'objectif de cette délimitation stricte était de limiter au maximum les problèmes d'alignement entre CP de différents annotateurs dans la suite. Lors

de l'annotation d'un **cp_axio**, l'annotateur pouvait renseigner les propriétés suivantes, systématiquement associées aux unités **cp_axio** dans le modèle d'annotation fourni avec l'outil Glozz :

- **polarity** : '+' ou '-' pour rendre compte de la polarité de l'opinion exprimée, s'il y en a une ;
- **focus** : 's' ou 'n' ou 'b' pour "sharpen", "neutral" et "blur" - cette propriété rend compte de la « prototypicalité » ;
- **force** : 'h' ou 'm' ou 'l' pour "high', "medium" et "low" - cette propriété rend compte de la force ou « intensité » de l'opinion exprimée

Des exemples illustraient les notions de focus et de force, inspirées de (Martin et White, 2005).

Il était aussi demandé aux experts d'annoter si possible la cible d'un CP. A priori réalisable pour tous les CP, cp_axio ou cp_qua, cette annotation se réalisait en 2 étapes : annoter l'unité dans la phrase qui est la cible du CP (exception : dans la phrase précédente ou suivante pour les CP au sein de phrases averbales) en utilisant le type de segment "tgt_ent" ("target entity") fourni dans le modèle d'annotation, puis mettre en relation le CP et sa cible en utilisant la relation "opi_tgt" ("opinion target"). Là encore, des exemples étaient fournis pour illustrer le résultat attendu.

	ex phrase
groupe adjectival	***Méthodique, rationnel,*** *il aime la gestion quotidienne de l'entreprise, le métier, mais n'entend rien à la stratégie et au long terme.*
construction absolue	*<Titre > Pierre Barbéris,* ***une main de fer et un rire de velours****.*
participe	*Discret, voire solitaire,* ***réputé pour son sang froid,*** *il n'a pas été surnommé pour rien « le Boa » après avoir avalé le Crédit Lyonnais et son amour de la corrida ne l'a pas transformé en victime consentante.*
adverbial	***Habilement****, le Royaume-Uni a de son côté négocié une dérogation avantageuse : ...*
circonstant prépositionnel	***En position toujours délicate,*** *Citi affiche pour sa part une perte nette de 2,8 milliards sur les trois dernières années.*
GN sans déterminant	***Femme de caractère****, cette pasionaria de la pub cumule les titres chez Havas.*
GN avec déterminant	*Olga, la serveuse,* ***une femme à poigne****, a bien du mal à faire régner l'ordre.*

Tableau 1 : Extrait de l'exemplification fournie aux annotateurs

4.1.2 LA PLATE-FORME GLOZZ

Présentation générale

Dans la continuité d'un grand nombre de travaux sur corpus, en linguistique et en TAL, notre travail impose l'établissement d'annotations manuelles permettant l'exploration du phénomène étudié et auxquelles des traitements automatiques pourront ultérieurement être confrontés. Si

la mise en place de telles annotations peut être envisagée de manière *ad hoc*, sur la base d'outils et de modèles d'annotations retenus pour cette seule occasion, il est toutefois évidemment préférable d'exploiter des procédures, outils et formats aussi génériques que possible, pour limiter le coût de mise en œuvre des campagnes d'annotation, pour assurer la capitalisation efficace des résultats, pour permettre de les exploiter dans d'autres perspectives scientifiques, pour rendre possible leur diffusion au sein de la communauté...

Pour le travail présenté dans cet article, nous avons utilisé la plate-forme Glozz (http://www.glozz.org) (Widlöcher et Mathet, 2009), développée au sein du GREYC. Répondant aux contraintes de généricité évoquées ci-dessus, elle propose un environnement d'exploration de corpus et d'annotation fortement configurable et non limité a priori au contexte discursif dans lequel elle a initialement vu le jour, dans le cadre du projet Annodis (Péry-Woodley *et al.*, 2009). Glozz repose sur un méta-modèle générique qui permet l'annotation de structures linguistiques variées, observables à différents niveaux de granularité, qui reposent sur des segments, des relations ou des dispositifs plus complexes, nommés schémas. Des caractérisations peuvent être associées aux objets annotées, sous forme de structures de traits. Pour une campagne d'annotation donnée, un modèle d'annotation doit être défini, qui spécifie les objets linguistiques disponibles, et le format des caractérisations dont ils devront faire l'objet. Guidé et contraint par ce modèle, l'annotateur produit graphiquement de nouvelles annotations, en les localisant dans le texte et en y associant les indications attendues.

Dans ce but, il peut s'appuyer sur les différentes informations portées par le corpus, qu'il s'agisse du contenu textuel initial ou d'informations linguistiques (morphologiques, syntaxiques, sémantiques...) résultant éventuellement d'une annotation préalable, manuelle ou automatique. Glozz simplifie l'exploitation de ces indications, par le biais de différents outils de navigation. Il offre notamment une vue synthétique du corpus donnant un accès rapide aux zones porteuses d'indices, permet l'emphase graphique des objets linguistiques pertinents et le filtrage des objets insignifiants pour une tâche donnée. De plus, Glozz intègre des outils de recherche avancée pour guider l'exploration du corpus. L'utilisateur peut notamment effectuer une recherche plein-texte, restreindre une telle recherche à des contextes spécifiques correspondant à des environnements linguistiques particuliers, ou naviguer rapidement entre les différentes instances d'un même type d'objet linguistique, en exprimant éventuellement des contraintes sur les traits portés par les objets. Disposant d'un pouvoir expressif plus élevé, le langage d'interrogation GlozzQL permet la description de classes d'objets vérifiant certaines contraintes (contraintes sur le contexte d'apparition, sur la présence de traits, sur les relations entretenues...) et la navigation entre les instances de ces classes. Ainsi, toute annotation peut s'inscrire dans un processus d'enrichissement incrémental : en s'appuyant sur des informations préexistantes, l'annotateur produit de nouveaux objets, qui pourront eux-mêmes, à leur tour, devenir "indices", lors de l'annotation d'éléments d'ordre supérieur.

Enfin, Glozz propose différents outils dédiés à l'évaluation des annotations réalisées selon son méta-modèle. Il permet notamment d'aligner différentes annotations réalisées sur un même corpus et de produire une synthèse de cet alignement, synthèse indiquant par exemple l'accord entre annotateurs, les distances entre annotations...

Utilisation de Glozz pour notre étude

Dans le cadre du présent travail, la plate-forme Glozz a été utilisée à l'occasion des phases suivantes :

1. Adaptation de l'application à la campagne d'annotation initiale. À ce niveau, nous avons tout d'abord défini le modèle d'annotation indiquant les unités et relations utilisables par

l'annotateur, et spécifiant le format des représentations symboliques devant y être associées. Nous avons d'autre part précisé, par le biais d'une feuille de style, la manière dont ces objets devaient être visuellement représentés dans l'interface d'annotation. Enfin, nous avons assuré la conversion du corpus dans un format exploitable dans Glozz.

1) Annotation manuelle libre. Contraints par le modèle d'annotation préalablement défini, les annotateurs ont alors procédé à la première phase d'annotation, consistant à délimiter les unités (cp_axio, cp_qua et tgt_ent), à indiquer les relations (opi_tgt) entre CP et cibles, puis à renseigner, pour ces différents éléments, les structures de traits spécifiées par le modèle (polarity, force, focus...). Ces différentes tâches ont été réalisées graphiquement, dans un environnement conforme à celui que représente la figure , qui sera commentée ci-après.

1. Alignement des annotations. À l'issue de cette phase d'annotation, la plate-forme a été utilisée pour procéder à l'alignement des annotations produites par les différents annotateurs, et pour produire une synthèse indiquant l'accord entre annotateurs, la distance entre les bornes droites des segments délimités, les valeurs des traits associés par les différents annotateurs... Cette synthèse a été utilisée pour réaliser la première série de mesures.

    - Préparation de l'annotation contrôlée. À l'issue de l'alignement décrit ci-dessus, Glozz a également été utilisé pour récupérer l'union des segments annotés par chacun des annotateurs. Quand les annotateurs avaient indiqué une borne droite différente, nous avons retenu la portée maximale. Un nouveau modèle d'annotation a par ailleurs été mis en place, disposant simplement des types *accept* et *reject*, modèle devant permettre, lors de l'annotation contrôlée, de distinguer les CP à valeur effectivement axiologique.

1. Annotation contrôlée. Chargées dans Glozz, les annotations résultant de cette union ont ensuite été éditées à la lumière de ce nouveau modèle d'annotation, afin que chaque annotateur puisse simplement se prononcer sur le caractère acceptable, ou non, de chacune de ces annotations.

A) Enfin, Glozz a permis de produire la synthèse de ces annotations contrôlées, synthèse à partir de laquelle a été effectuée la seconde série de mesures.

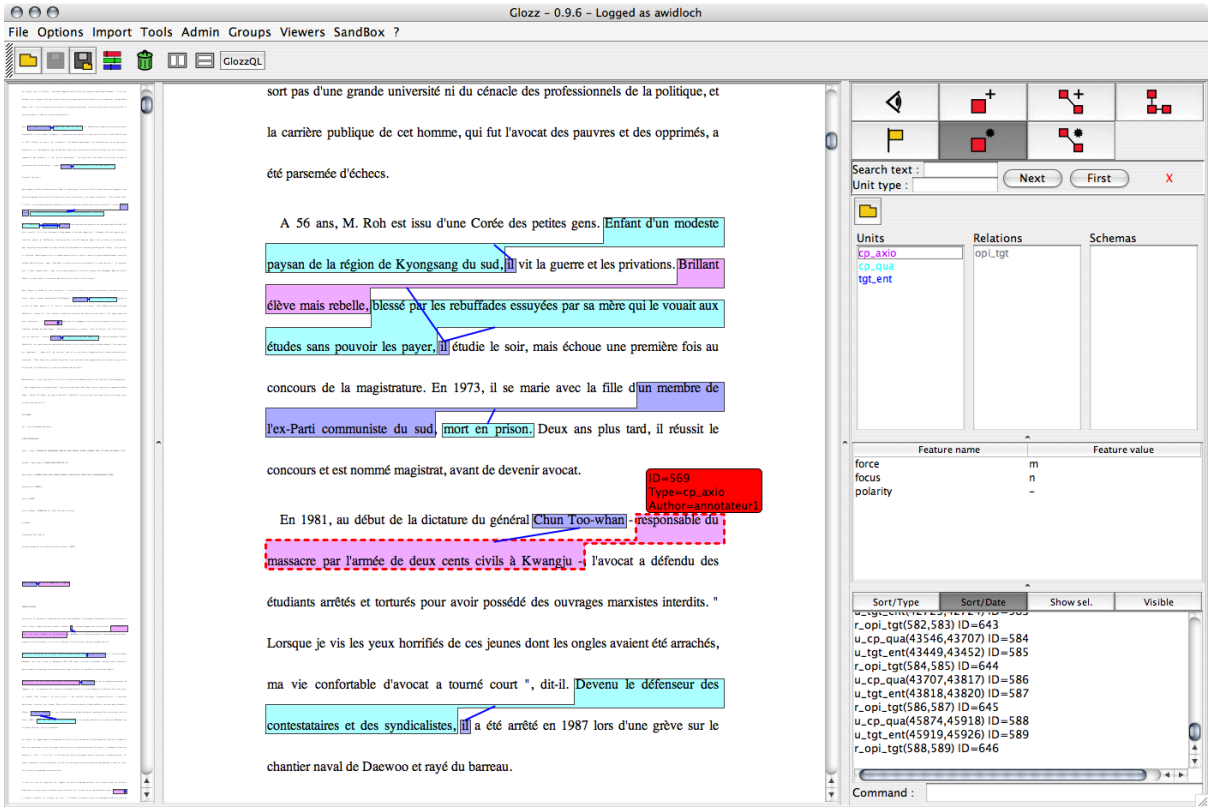


Figure 1 : Interface principale de Glozz

La figure 1 représente l'interface principale de Glozz, lors de la phase d'annotation manuelle libre. La colonne de gauche présente une vue synthétique du document permettant de naviguer rapidement entre ses parties. La partie centrale constitue la zone de travail principale, au sein de laquelle l'utilisateur peut manipuler graphiquement ses annotations. La partie droite de l'interface permet notamment à l'annotateur de sélectionner le type linguistique des objets manipulés et d'éditer les structures de traits qui y sont associées.

4.2 Premiers résultats

4.2.1 ANNOTATION LIBRE DE CP AXIOLOGIQUES OU NON

Le protocole a été proposé à quatre experts, ainsi qu'une réunion informelle permettant d'échanger sur les différents types de CP déjà observés au préalable afin de familiariser chacun aux unités à repérer. Le procédé automatique d'alignement retenu s'appuye sur la particularité des annotations attendues : des segments fortement délimités par la ponctuation, une distance quasi nulle entre deux annotations différentes. Avec cette méthode, nous obtenons 181 segments annotés, dont 32 en commun aux 4 annotateurs (18%), 33 communs à 3 (18%), 26 à 2 (14%) et 90 (50%) annotés seulement par un expert.

Sur cette base, nous avons calculé deux taux d'accord distincts, l'un pour quantifier l'accord sur la notion de CP, l'autre pour mesurer plus spécifiquement l'accord sur les CP axiologiques, cœur de notre étude. Nous avons utilisé le Kappa de Fleiss pour mesurer dans les deux cas un taux global aux quatre annotateurs, puis le Kappa de Cohen pour observer les taux d'accords entre annotateurs deux à deux.

La première série de mesures porte sur le simple fait d'avoir annoté un même segment CP, sans prendre en considération la nature de l'annotation (**cp_axio** ou **cp_qua** dans le protocole précédent) : l'accord entre les 4 vaut 0.13 – mauvais ; les 6 accords deux à deux sont -0.1 (très mauvais), 0.12, 0.13, 0.16, 0.2 et 0.36 (médiocre). La deuxième série ne concerne plus que les CP axiologiques : l'accord entre les 4 vaut -0.06 – très mauvais ; les 6 accords deux à deux sont 0.15 (mauvais), 0.24, 0.3, 0.33, 0.4 et 0.53 (modéré).

Devant la multiplicité des tâches proposées, les différents experts se sont visiblement focalisé chacun sur des objectifs différents, certains renseignant les structures de traits, d'autres repérant les cibles, et leur lecture des textes semble en avoir été grandement influencée. Enfin, la frontière entre ce qui est axiologique et ce qui ne l'est pas est souvent difficile à établir ; les annotateurs ont rencontré certaines difficultés à décider de ce qui relevait effectivement du langage de l'évaluation. Ce point a déjà été noté dans des travaux précédents : *« Passer de l'énumération des propriétés objectives d'un objet à son évaluation axiologique, c'est effectuer, toujours, en prenant appui sur ses compétences culturelle et idéologique, un certain 'saut interprétatif' (plus ou moins audacieux, plus ou moins contestable) »* (Kerbrat-Orecchioni, 2002, p.107). Les résultats obtenus ne permettent pas d'envisager en l'état la mise en place d'un corpus de référence. Au contraire, ils semblent indiquer que la tâche est subjective en soi, menant plus au désaccord qu'à l'accord... Cependant, les taux deux à deux concernant la nature axiologique des segments annotés laissent penser qu'un accord reste possible, à condition de simplifier la tâche d'annotation. Nous avons donc décidé pour la suite de la campagne de veiller à ne pas solliciter l'attention des annotateurs sur de multiples problèmes mais au contraire de la focaliser sur un unique objectif.

4.2.2 ANNOTATION CONTRÔLÉE DE SEGMENTS

Pour la deuxième phase d'annotation, l'ensemble des 191 segments repérés à l'issue de la première phase ont été soumis à 4 experts, 3 ayant déjà participé à la première phase. Chaque expert avait désormais pour consigne de préciser uniquement si le segment était un CP axio ou non. Dans la pratique, la plate-forme Glozz présentait par défaut les 191 segments comme CP axio, et l'annotateur devait en changer le type s'il jugeait cette information fausse.

Le taux d'accord entre les 4 vaut désormais 0.72 – bon ; les 6 taux deux à deux sont 0.56 (modéré), 0.58, 0.64, 0.84, 0.86 et 0.87 (excellent). Les 3 taux deux à deux les plus faibles impliquent un même annotateur A0, ayant participé aux deux phases de la campagne. L'analyse qualitative des résultats laisse certes paraître quelques erreurs d'inattention ou possiblement dues à un mauvais usage de l'interface., mais elles ne suffisent pas à expliquer les différences entre l'annotateur A0 et les autres.

Au vu de ces nouveaux résultats, la poursuite de la campagne peut de nouveau être envisagée. Cependant, vu le coût de la mise en place de ce type de campagne, il convient de faire encore quelques tests pour trouver une méthode permettant, après mesure d'accord satisfaisante, de laisser les experts annoter chacun des textes différents afin d'en obtenir une plus grande quantité pour mener les évaluations des outils informatiques. Or, les résultats obtenus sont le fruit d'une annotation multiple au départ. Nous envisageons de fait encore deux approches à tester :

- une annotation automatique des segments entre ponctuations pour présenter aux experts qui décideront simplement de leur caractère CP axiologique ou non.
- une annotation par les experts des segments qu'ils jugent CP axiologiques, sur texte libre.

La première méthode est séduisante, car elle permet de se rapprocher de la manière dont se sont déroulées les deux phases présentées ici. Cependant, nous avons l'ambition de traiter

aussi d'autres formes que les constituants périphériques, et nous n'avons pas l'assurance de pouvoir annoter automatiquement ces nouvelles formes à venir. C'est pourquoi nous préférons a priori la deuxième méthode, si elle donne de bons résultats, car elle peut être reproduite pour d'autres types d'énoncés.

Cependant, il convient aussi de noter soit le coût élevé soit la portée restreinte de l'annotation envisageable. Nous espérions pouvoir mettre en place une annotation complexe, faisant apparaître les propriétés des énoncés évaluatifs. Les résultats et les méthodes possibles pour poursuivre laissent peu de place à cette ambition : il faudrait envisager, si l'on poursuit à l'extrême le même raisonnement, une campagne d'annotation pour chaque propriété étudiée... Une réflexion est menée actuellement sur ce problème, afin d'optimiser le coût de la mise en place d'un corpus de référence.

5 CONCLUSION, PERSPECTIVES

Nous avons présenté une étude de l'expression de jugements d'opinion au sein de constituants détachés. Une analyse du phénomène sur un premier corpus constitué essentiellement de « Portraits » issus du journal économique Les Échos a permis de proposer une vingtaine de patrons récurrents. Nous avons alors mis en place une expérimentation informatique dans le but de tester ces patrons sur d'autres ressources. La mise en œuvre a été réalisée sous forme de règles d'extractions dans une chaîne d'analyse Linguastream pour 10 des patrons observés. Testées sur un nouveau corpus constitué de portraits issus du journal généraliste Le Monde, ces règles ont permis de montrer que les patrons continuent d'être utilisés pour exprimer des jugements dans ce type de textes, mais avec des emplois qui se distinguent de ceux observés initialement : ces jugements peuvent désormais porter sur des objets alors que notre première étude se limitait à ceux portés sur des personnes et sur leurs actions. De nouveaux phénomènes ont pu être mis en évidence par cette expérimentation, notamment l'existence de nombreuses énumérations de constituants détachés concentrés dans quelques textes du corpus, l'usage de connecteurs au sein des constituants de nature adjectivale ainsi que l'utilisation récurrente de phrases averbales ; phénomènes qui sont autant de pistes pour de futurs travaux.

Dans une dernière partie, nous avons envisagé la constitution d'un corpus de référence pour permettre une évaluation quantifiée des résultats obtenus. Nous avons dans un premier temps opté pour une campagne d'annotation qui prenne d'emblée en considération la complexité du phénomène étudié et renseigne différentes propriétés des opinions selon un protocole inspiré du modèle de l'Appraisal (Martin et White, 2005). Les taux d'accord mesurés entre annotateurs, très mauvais, nous ont conduit à réviser cette première approche pour concentrer l'effort d'annotation sur le simple repérage des segments textuels relevant du phénomène étudié. Les taux d'accord obtenus sont alors nettement meilleurs, permettant d'envisager de poursuivre la mise en place d'un corpus de référence, mais posant cependant de nouveaux problèmes. Le protocole suivi est en effet adapté à une annotation très simple qui ne rend pas compte de la complexité des phénomènes en jeu ; une réflexion nous paraît nécessaire pour envisager un enrichissement des structures annotées à moindre coût, mais avec des taux d'accord entre experts humains qui permettent effectivement de considérer le résultat obtenu comme pouvant servir de référence.

La poursuite de l'étude vise actuellement à enrichir le modèle initial, tant sur le plan des ressources que par l'obtention systématique de nouveaux observables. Si la mise en œuvre des patrons initiaux peut mener à un apprentissage de lexique de manière immédiate, la détection de nouveaux patrons pose des questions intéressantes en matière de fouille de données

textuelles auxquelles nous cherchons à apporter différents éléments de réponse. Plus précisément, nous explorerons deux types d'approches. La première tire parti d'une expérience qui a été menée dans un domaine très différent, celui des textes biologiques et génétiques. Ce travail a montré l'intérêt d'utiliser les motifs séquentiels pour apprendre automatiquement des patrons linguistiques en vue de réaliser le repérage et l'extraction de relations sémantiques entre entités nommées (Plantevit et Charnois, 2009). Nous pensons utiliser cette approche pour acquérir de nouveaux patrons, à partir de corpus non annotés, et en intégrant diverses contraintes, notamment de type linguistique, pour sélectionner les motifs intéressants. La deuxième approche que nous voulons mettre en œuvre utilisera, sur le corpus annoté, un nouveau type de motifs, les motifs LSR, qui combinent les avantages des motifs séquentiels et des motifs ensemblistes (Plantevit et al, 2009), afin d'explorer le contexte des expressions évaluatives. Cette exploration permettra en phase d'apprentissage une meilleure sélection des patrons et, en phase de tests, l'utilisation de critères contextuels lors de l'application des patrons.

6 REMERCIEMENTS

La plate-forme Glozz a été réalisée dans le cadre du projet ANNODIS (Annotation Discursive, un corpus de référence annoté discursivement, outils d'annotation et d'exploitation de corpus annotés), bénéficiant d'une aide de l'ANR, Agence Nationale de la Recherche.

Les travaux que nous menons sur le discours évaluatif se poursuivent quant à eux dans le cadre du projet ONTOPITEX (Modèles linguistiques et ontologies – Extraction informatique et caractérisation d'opinions et de jugements d'évaluation dans les textes), bénéficiant d'une aide de l'ANR portant la référence ANR-08-CORD-009.

7 RÉFÉRENCES

Baroni M. et Vegnaduzzo S. (2004). « Identifying subjective adjectives through web-based mutual information ». Dans *Proceedings of KONVENS-04*. Vienna, Austria. p. 17-24.

Bilhaut F. et Widlöcher A. (2006). « LinguaStream: An Integrated Environment for Computational Linguistics Experimentation ». Dans *Proceedings of EACL 2006, the 11th Conference of the European Chapter of the Association of Computational Linguistics*. p. 95-98.

Combettes B. (1998). *Constructions détachées en français.* Ophrys, Collection L'essentiel Français.

Esuli A. et Sebastiani F. (2006).« SentiWordNet : A Publicly Available Lexical Resource for Opinion Mining ». Dans *Proceedings of LREC-06, the 5th Conference on Language ressources and Evaluation*. Genova, Italy.

Ferrari S., Charnois T., Mathet Y., Rioult F. et Legallois D. (2009). « Analyse de discours évaluatif, modèle linguistique et applications ». Dans *RNTI, Revue des Nouvelles Technologies de l'Information, vol E-17, numéro spécial Fouille des Données d'Opinions.* p. 71-94.

Grouin C., Hureau-Plantet M., Paroubek P. et Berthelin J.-B. (2009). « DEFT'07 : une campagne d'évaluation en fouille d'opinion ». Dans *RNTI, Revue des Nouvelles Technologies de l'Information, vol E-17, numéro spécial Fouille des Données d'Opinions*. p. 1-24

Guimier C. (1996). *Les adverbes du français*. Ophrys Collection L'essentiel Français.

Hatzivassiloglou V. et McKeown K. (1997). « Predicting the semantic orientation of adjectives ». Dans *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain : ACL. p. 174-181.

Hu M., Liu B. (2004). « Mining Opinion Features in Customer Reviews ». Dans *Proceedings of Nineteeth National Conference on Artificial Intelligence (AAAI-2004)*.

Jackiewicz A. (2009). « L'évaluation à la périphérie du prédicat : constructions, lexiques et relations sémantiques ». *Arena Romanistica, N°4, Actes du 28e Colloque international sur le Lexique et la Grammaire.* Bergen, Norvège.

Jackiewicz A., Charnois T. et Ferrari S. (2009). « Jugements d'évaluation et constituants périphériques ». Dans *Actes de TALN09, 16e conférence sur le traitement automatique des langues naturelles*. France : Senlis.

Kerbrat-Orecchioni C. (2002). *L'énonciation – de la subjectivité dans le langage*. 4e édition. France : Armand Colin.

Martin J. R. et White P. R. R. (2005). *The Language of Evaluation, Appraisal in English*. London, New York : Palgrave Macmillan.

Pang B. et Lee L. (2004). « A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts ». Dans *Proceedings of the 42nd Annual Meeting of the ACL*.

Pang B. et Lee L. (2008). *Opinion mining and sentiment analysis*. Now Publishers.

Péry-Woodley M.-P., Asher N., Enjalbert P., Benamara F., Bras M., Fabre C., Ferrari S., Ho-Dac L.-M., Le Draoulec A., Mathet Y., Muller P., Prévot L., Rebeyrolle J., Tanguy L., Vergez-Couret M., Vieu L. et Widlöcher A. (2009). « ANNODIS : une approche outillée de l'annotation de structures discursives ». Dans *Actes de TALN09, 16e conférence sur le traitement automatique des langues naturelles*. France : Senlis.

Plantevit M., Charnois T., Kléma J. Rigotti C. et Crémilleux B. (2009). « Combining sequence and itemset mining to discover named entities in biomedical texts: a new type of pattern ». *International Journal of Data Mining, Modelling and Management*, vol. 1, n°2. p. 119-148.

Plantevit M. et Charnois T. (2009). « Motifs séquentiels pour l'extraction d'information : illustration sur le problème de la détection d'interactions entre gènes ». Dans *Actes de TALN09, 16e conférence sur le traitement automatique des langues naturelles*. France : Senlis.

Read J., Hope D. et Carroll J. (2007). « Annotating Expressions of Appraisal in English ». Dans *Proceedings of the Linguistic Annotation Workshop*. ACL.

Riloff E. et Wiebe J. (2003). « Learning extraction patterns for subjective expressions ». Dans *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*. Sapporo, Japan : ACL. p. 105-112.

Schmid H. (1994). « Probabilistic Part-off Speech Tagging Using Decision Trees ». Dans *Proceedings of the First International Conference on New Methods in Natural Language Processing (NemLap-94)*. p. 44-49.

Turney P. D. (2002). « Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews ». Dans *Proceedings of the 40th Annual Meeting of the ACL*.

Turney P. et Littman M. L. (2003). « Measuring praise and criticism : Inference of semantic orientation from association ». *ACM Transactions on Information Systems (TOIS)* 21(4), p. 315-346.

Vernier M. et Ferrari S. (2007). « Tracking evaluation in discourse ». Dans *Proceedings of ASOS07, Workshop on Applications of Semantics, Opinions and Sentiments*.

Whitelaw C., Garg N. et Argamon S. (2005). « Using Appraisal taxonomies for sentiment analysis ». Dans *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*.

Widlöcher A. et Bilhaut F. (2008). « Articulation des traitements en TAL - Principes méthodologiques et mise en œuvre dans la plate-forme LinguaStream ». *Revue Traitement Automatique des Langues (TAL)*, 49(2), p. 73-101.

Widlöcher A. et Mathet Y. (2009). « La plate-forme Glozz : environnement d'annotation et d'exploration de corpus ». Dans *Actes de TALN09, 16e conférence sur le traitement automatique des langues naturelles*. France : Senlis.

Wiebe J., Wilson T. et Cardie C. (2005). « Annotating Expressions of Opinions and Emotions ». *Language. Language Resources and Evaluation* 39, issue 2-3.

Wiebe J. et Mihalcea R. (2006). « Word sense and subjectivity ». Dans *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sidney, Australia : ACL.

Yu H. et Hatzivassiloglou V. (2003). *Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. Dans Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*.

STRATÉGIE DE CONSTITUTION D'UN CORPUS DE TEXTES SCOLAIRES DÉDIÉ À DES ÉTUDES MÉTALEXICOGRAPHIQUES ET À LA CONCEPTION D'UN MODULE D'HYPERAPPEL DE DICTIONNAIRE

Nathalie Gasiglia et Stavroula Markezi
Université Charles-de-Gaulle – Lille 3
STL – U.M.R. 8163 du C.N.R.S.

RÉSUMÉ

Le projet au sein duquel cette étude s'intègre fonde sur une analyse métalexicographique de dictionnaires existants des spéculations relatives à des dictionnaires électroniques qui pourraient être créés.

L'objectif métalexicographique principal de ce projet consiste à étudier comment un corpus de textes issus de manuels scolaires peut permettre d'évaluer les services susceptibles d'être effectivement rendus aux écoliers par des dictionnaires pour l'école élémentaire et d'imaginer de possibles améliorations de ces répertoires.

L'objectif applicatif, lui, est à l'interface du TAL et de l'informatique éditoriale : il s'agit de fonder sur une analyse du corpus la conception d'un module d'hyperappel d'un dictionnaire électronique conçu de manière telle qu'il orienterait mieux les usagers vers les descriptions lexicales pertinentes que ceux qui sont proposés actuellement.

La présente contribution se concentre sur des questions méthodologiques de constitution puis d'exploration du corpus : dans un premier temps, son élaboration, qui implique la sélection et la numérisation des documents primaires, leur étiquetage (avec introduction des lemmes et des codes morphosyntaxiques de chaque item relus et corrigés) et la constitution du corpus xmlisé, structuré en conformité avec les recommandations de la TEI ; dans un second temps, la mise en correspondance, chaque fois qu'elle est possible, d'une part de chaque item lexical de chaque texte avec un item utilisé comme adresse d'article dans l'un des dictionnaires scolaires pris en compte, et d'autre part des sens attestés dans le corpus avec ceux qui sont décrits dans le répertoire considéré.

Au terme de ces appariements (validés manuellement), les indications d'adressage introduites dans le balisage du corpus donnent les moyens d'évaluer la proportion d'items auxquels correspond une description dans le dictionnaire. Un retour critique sur les résultats obtenus permet de mettre en perspective les choix d'étiquetage et les résultats de la procédure de repérage des adresses et des subdivisions d'articles pertinentes.

1 INTRODUCTION

Le projet auquel se rattache cette contribution fonde sur une analyse métalexicographique de dictionnaires existants des spéculations relatives à des dictionnaires électroniques qui pourraient être créés.

L'objectif métalexicographique principal de ce projet consiste à étudier comment un corpus de textes issus de manuels scolaires peut permettre d'évaluer les services susceptibles d'être effectivement rendus aux écoliers par des dictionnaires pour l'école élémentaire et d'imaginer de possibles améliorations de ces répertoires, ce qui prolonge la réflexion de Gasiglia (2009 : 267-289 et à paraître a) en prenant appui sur la confrontation de contenus dictionnairiques et de productions textuelles destinés aux mêmes publics.

À cette visée analytique s'ajoute un objectif applicatif qui est à l'interface du TAL et de l'informatique éditoriale : il s'agit de la conception d'un module d'hyperappel d'un dictionnaire électronique (en imaginant que celui-ci pourrait être inclus dans un environnement numérique de travail scolaire qui réunirait par ailleurs des textes et exercices de français, de mathématiques ou de matières d'éveil) conçu de manière telle qu'il orienterait mieux les usagers vers les descriptions lexicales pertinentes que ceux qui sont proposés actuellement.

Complémentairement, le corpus constitué dans le cadre de ce projet diversifiera la documentation de différentes études métalexicographiques engagées antérieurement et relatives aux traitements, dans les dictionnaires destinés aux jeunes lecteurs, d'une part des emprunts (cf. Gasiglia (2008a, 2008b, 2008c)) et d'autre part des items de toutes origines dont les descriptions proposent des marquages pour certains emplois (cf. Corbin & Gasiglia (à paraître)), et en particulier ceux qui sont considérés comme relevant du "parler enfantin" (cf. Gasiglia (à paraître b)) ou ceux qui sont stigmatisés (cf. Gasiglia (à paraître c)). Ces études, qui reposaient jusqu'ici sur une documentation exclusivement dictionnairique, trouveront dans le corpus pédagogique constitué un point d'accès à des usages langagiers destinés aux jeunes lecteurs qui permettra d'étudier si les items empruntés ou marqués y sont employés et, dans l'affirmative, dans quels contextes ils le sont, ce qui pourra induire que leur étude cesse d'être strictement métalexicographique pour prendre aussi une dimension linguistique. Enfin, c'est également sur ce dernier terrain, mais en morphologie cette fois, que l'exploration du corpus fournira de précieuses informations permettant en particulier d'évaluer quels procédés de construction de mots sont à la source des mots construits qui y sont attestés et d'étudier si leur régularité peut avoir influé sur la décision de décrire ou non ces mots dans les dictionnaires scolaires, ce qui fera écho à Gasiglia (2008d).

La présente contribution se concentre sur des questions méthodologiques de constitution du corpus (les seules tâches effectivement réalisées par les deux rédactrices durant le stage de fin de master effectué par la seconde à l'UMR STL de mai à août 2009), mais, pour mettre celles-ci en perspective, elle évoque les incidences des choix opérés sur les deux premières finalités envisagées :

– l'appréciation des services rendus par les dictionnaires, concernant laquelle nous avons choisi de nous concentrer sur l'évaluation de l'aide à la compréhension, qui s'apprécie mieux à partir de l'exploration d'un corpus comme celui que nous avons constitué que l'aide à l'expression ;

– l'élaboration d'un module de consultation par hyperappel, dans laquelle nous distinguons deux mises en correspondance de complexité distincte : l'appariement de l'item en contexte d'appel avec un item en adresse dans le dictionnaire et celui de l'acception déterminable en contexte avec l'une des descriptions syntaxico-sémantiques qu'il propose.

2 SÉLECTION DES DOCUMENTS PRIMAIRES DU CORPUS

Les objectifs étant posés, nous pouvons exposer comment nous avons engagé la sélection des documents primaires qui constituent le corpus et ce qui a motivé que nous arrêtions notre choix sur les textes d'un manuel de français pour les élèves du CE1 (la troisième année du cycle 2, le cycle des apprentissages fondamentaux de l'école primaire) publié par Hachette, éditeur également présent sur le marché des dictionnaires scolaires. Ce manuel fait partie d'une collection – *À portée de mots* – constituée de quatre titres destinés respectivement aux élèves du CE1 et des CE2, CM1 et CM2 (les trois niveaux du cycle 3, le cycle des approfondissements), dont les trois premiers établissent un lien particulier entre leur section de vocabulaire et les deux dictionnaires de cet éditeur destinés aux élèves des cycles correspondants en dupliquant des pages de ces derniers comme supports pour les exercices d'initiation à la consultation dictionnairique. Le volume destiné aux élèves du CE1 (celui publié en 2003 comme la réédition révisée de 2009) propose ainsi trois pages de l'édition de 1996 du *Dictionnaire Hachette benjamin* ou de son retirage de 2002 (pages 98-99 du dictionnaire en pages 114-115 du manuel en 2003 et 130-131 en 2009, et page 318 en

page 118 en 2003 (cf. figure 1) et 134 en 2009), associées à trois pages d'exercices (pages 116-117 et 119 en 2003, 132-133 et 135 en 2009) portant sur les articles présents dans les pages reproduites ou sur ceux d'autres pages (par exemple, en page 119 du manuel de 2003, les exercices 1 et 3 portent sur les articles reproduits page 118 alors que les exercices 4 et 5 portent sur d'autres articles et peuvent être effectués avec n'importe quel dictionnaire (cf. figure 1)).

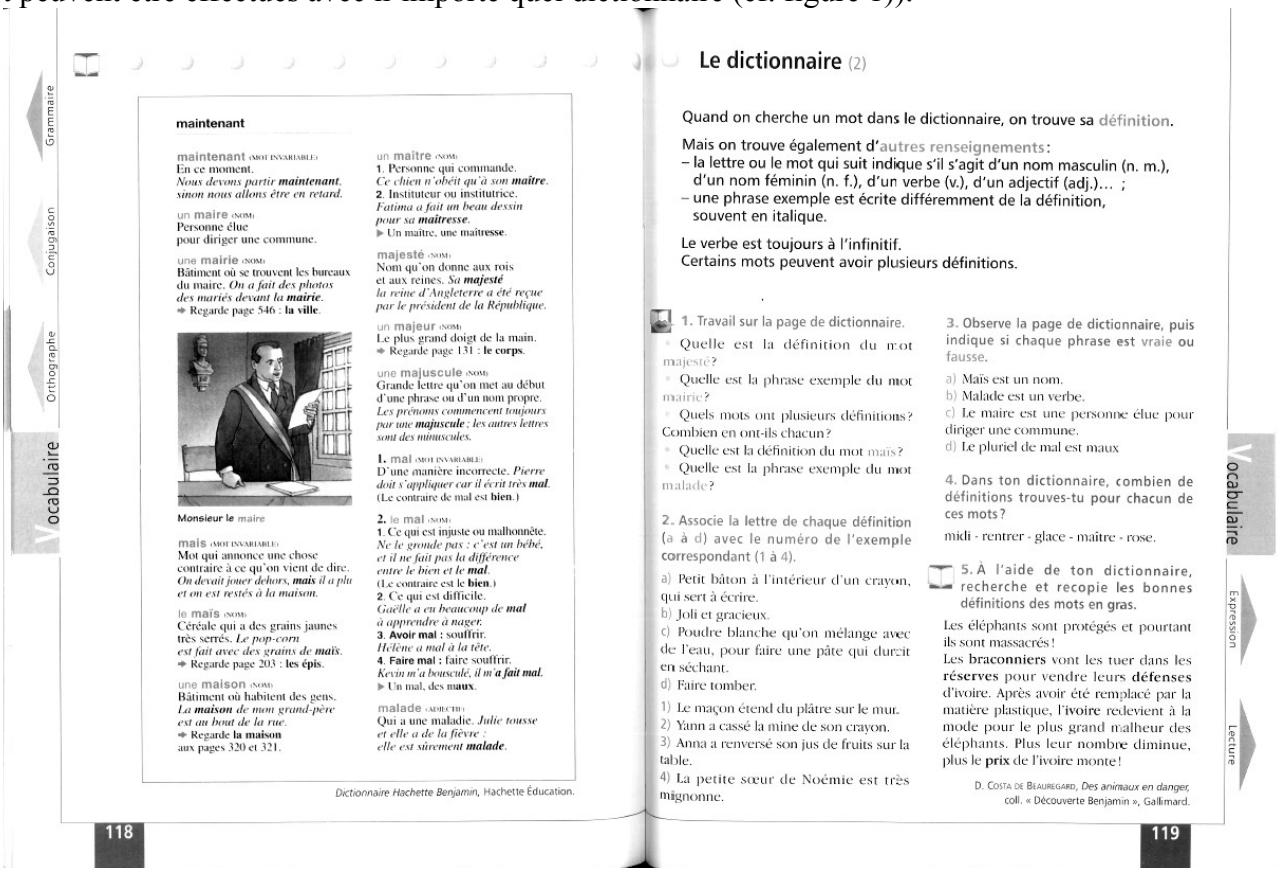


Figure 1 : Manuel À portée de mots CE1 de 2003, pages 118-119

Le volume destiné aux élèves du CE1 étant *a priori* celui dont les textes présentent le moins de difficultés de compréhension lexicale, il nous intéressait, d'un point de vue métalexicographique, de commencer par lui la comparaison des informations nécessaires pour éclairer l'interprétation d'un mot lu dans un manuel avec celles présentes dans le dictionnaire pour le cycle 2. Le projet se limite actuellement à ce répertoire mais, en imaginant qu'un élève de CE1 accéderait éventuellement à celui destiné à ses aînés, il mériterait d'être étendu au dictionnaire pour le cycle 3.

Une fois la sélection du manuel source établie, celle des textes retenus en son sein s'est faite en fonction de leur longueur relative : les textes narratifs (cf. figure 2) ou documentaires et occasionnellement les poèmes qui introduisent chaque séquence pédagogique semblaient pouvoir être plus porteurs d'indices permettant aux élèves ou à un futur système d'hyperappel d'un dictionnaire de choisir dans les articles de celui-ci les descriptions de sens pertinentes en fonction des contextes sources que les textes des exercices qui, étant fractionnés, s'y prêtent moins. La pertinence du choix ne tient pas seulement à cette commodité technique mais repose aussi sur le genre textuel impliqué, la "littérature enfantine", telle qu'elle s'offre aux enfants indépendamment du manuel (même si quelques textes émanent d'un des auteurs de celui-ci, une part importante puise dans des collections notoires et de large diffusion) et en tant qu'elle contribue à leur éducation en leur permettant de diversifier leurs connaissances au-delà des limites de leur expérience personnelle, ce qui n'est pas le cas de la métalangue du manuel (par ailleurs digne aussi d'intérêt, mais dont la connexion avec la consultation du dictionnaire n'a pas la même portée).

Grammaire

Ma sœur est une sorcière !

Ma sœur est une sorcière. On ne sait pas comment c'est arrivé. Maman dit que c'est à cause des livres Harry Potter que papa achète. Papa dit que c'est à cause de la mère de maman. Moi, je crois que ma sœur regarde trop la télévision. En plus, elle n'aime que les films fantastiques. Alors, forcément, il fallait s'y attendre. Mais, c'est bien embêtant…

L'autre jour, par exemple, j'avais emprunté sa console de jeux sans lui demander la permission. Quand elle est rentrée de son cours de potion magique, elle m'a regardé droit dans les yeux et a crié : « Carabistouille et fleur de nouille, par toutes les étoiles de l'univers, ce soir, tu diras tout de travers ! »

Eh bien, de toute la soirée, je n'ai pas pu me faire comprendre. C'était affreux. Quand maman m'a demandé :

– Théo, je te redonne de la soupe ?

J'ai répondu :

– Soupe ça maman la. N'aime non je pas.

Et hop, une assiette de soupe en plus…

Quand maman m'a proposé :

– Théo, tu veux de la crème au chocolat ?

J'ai dit :

– Plaît s'il oui, te maman oh.

Et la crème au chocolat, je n'ai fait que la regarder me passer sous le nez…

C'est terrible d'avoir une sœur sorcière !

J.-C. Lucas

Questions

a. D'après Théo, pourquoi sa sœur est-elle devenue sorcière ?
b. Est-ce que sa maman est du même avis ?
c. Quelle est la formule magique prononcée par la sœur de Théo ?
d. Quand sa maman lui a proposé de la crème au chocolat, que voulait répondre Théo ?

8

Figure 2 : Manuel À portée de mots CE1 de 2009, page 8
(texte déjà présent en 2003 mais signé en 2009)

3 DE LA NUMÉRISATION DES TEXTES À LA STRUCTURATION DU CORPUS

L'élaboration du corpus d'étude implique la numérisation des textes retenus. Cette tâche demande un travail minutieux mais elle ne pose pas de problème méthodologique particulier dans la mesure où nous cherchons à obtenir des fac-similés des originaux afin de les enregistrer dans un corpus XML structuré respectueux des recommandations de la *TEI*. Les choix relatifs à l'annotation des textes numérisés présentent, eux, plus d'enjeux.

Le but étant de mettre en correspondance les items lexicaux de chaque texte du manuel avec les entrées des dictionnaires pour le cycle 2 puis éventuellement pour le cycle 3 du même éditeur, nous avons commencé par faire étiqueter chaque texte par Cordial Analyseur, de façon à associer à chaque mot simple un lemme et un code morphosyntaxique. Le choix de Cordial Analyseur est motivé par le fait qu'il est développé par un éditeur d'outils de correction orthographique et grammaticale et qu'il nous semble raisonnable d'imaginer que, dans un environnement numérique de travail offert à des écoliers, il pourrait être jugé intéressant de fournir aux élèves un correcteur orthogrammatical et un hyperappel de leur dictionnaire, ce qui pourrait se faire en développant le second à partir des analyses que le premier est capable d'établir pour les textes en cours de rédaction et qu'il pourrait proposer pour les textes lus. Le recours à l'outil développé par Synapse Développement plutôt qu'à ceux des Éditions Diagonal (Prolexis) ou de Druide Informatique (Antidote) est motivé, lui, par le fait qu'il est à notre connaissance le seul à donner accès aux annotations produites par son module de lemmatisation et d'analyse morphosyntaxique (ce qui induit la commercialisation de Cordial Analyseur en parallèle du correcteur Cordial).

Les étiquetages qui ont été produits par Cordial Analyseur concernent les mots simples que cet outil sait délimiter et identifier. Nous nous sommes interrogées sur l'opportunité qu'il y avait à lui refaire étiqueter les textes en sélectionnant cette fois le repérage de toutes les unités linguistiques

polylexicales (mots composés, locutions, expressions) et à les faire parallèlement repérer par Unitex, ce qui nous aurait permis de croiser les produits des détections de ces deux outils pour établir une sélection large de celles qui sont attestées dans le corpus. Nous n'avons finalement pas retenu cette approche à cette étape de la recherche, en raison de la lourdeur de sa mise en œuvre. Dans les limites du temps de collaboration dont nous disposions (4 mois de stage), nous n'avons pas non plus engagé le réétiquetage par Cordial Analyseur seul, ce qui a impliqué que nous délaissions temporairement une partie de l'objectif métalexicographique (l'identification des expressions du corpus que le dictionnaire ne couvre pas) au bénéfice de la visée applicative orientée vers l'établissement des hyperappels, en limitant la détection des expressions du corpus à celles qui sont à la nomenclature du dictionnaire : lors de la mise en correspondance du texte lu et des articles de celui-ci, le script d'appariement des lemmes prévoit que, pour chaque item simple détecté et décrit dans le dictionnaire, si l'article le décrivant comporte la mention d'une adresse subordonnée polylexicale, la recherche de celle-ci est entreprise dans le contexte de l'occurrence en corpus (cf. § 4. points 3), 4) et 5)).

Après l'étiquetage de chaque texte par Cordial Analyseur, les produits de cet outil sont présentés de manière telle que les phrases sont délimitées et que chaque item figure sur une ligne, séparé par des tabulations de son lemme, de son code morphosyntaxique et d'un code d'identification mentionnant l'année de première édition du texte (2003 ou 2009), sa page, le rang de la phrase en son sein et celui de l'item, ce code étant généré par posttraitement à partir des numéros des phrases et des mots déterminés par l'analyseur et de l'année d'édition et la pagination mentionnées dans le nom du fichier contenant chacun des textes.

Ces produits de l'analyseur ont ensuite été corrigés manuellement. Les corrections apportées concernent en particulier les noms féminins comme *sœur*, qui est décrit dans les dictionnaires sous l'adresse **sœur** mais pour lequel le lemme du nom équivalent au masculin, *frère*, a été introduit lors de l'étiquetage, ce que montre la figure 3, où apparaissent dans la première colonne les délimitations de phrases et les codes d'identification des mots et ponctuations, dans la deuxième les items du corpus, dans les troisième et quatrième les lemmes et codes morphosyntaxiques proposés par Cordial Analyseur et validés ou ceux proposés au titre des corrections et, pour les items dont les annotations ont été corrigées, dans les cinquième et sixième les étiquetages initialement introduits et ayant donné lieu à correction. Cette mention des fautes corrigées manuellement permet *a posteriori* de détecter les fautes systématiques et de prévoir de les corriger automatiquement lors de l'annotation de futurs textes.

==== DEBUT DE PHRASE ====					
2003p008ph1-1	Ma	mon	DETPOSS		
2003p008ph1-2	sœur	sœur	NCFS	frère	NCFS
2003p008ph1-3	est	être	VINDP3S		
2003p008ph1-4	une	un	DETIFS		
2003p008ph1-5	sorcière	sorcier	NCFS		
2003p008ph1-6	!	!	PCTFORTE		
==== FIN DE PHRASE ====					
==== DEBUT DE PHRASE ====					
2003p008ph2-1	Ma	mon	DETPOSS		
2003p008ph2-2	sœur	sœur	NCFS	frère	NCFS
2003p008ph2-3	est	être	VINDP3S		
2003p008ph2-4	une	un	DETIFS		
2003p008ph2-5	sorcière	sorcier	NCFS		
2003p008ph2-6	.	.	PCTFORTE		
==== FIN DE PHRASE ====					
==== DEBUT DE PHRASE ====					
2003p008ph3-1	On	on	PPER3S		
2003p008ph3-2	ne	ne	ADV		
2003p008ph3-3	sait	savoir	VINDP3S		
2003p008ph3-4	pas	pas	ADV		
2003p008ph3-5	comment	comment	SUB		
2003p008ph3-6	c'	ce	PDS		
2003p008ph3-7	est	être	VINDP3S		
2003p008ph3-8	arrivé	arriver	VPARPMS		
2003p008ph3-9	.	.	PCTFORTE		
==== FIN DE PHRASE ====					

Figure 3 : Produit de Cordial Analyseur (relatif au texte de la page 8 du manuel À portée de mots CE1 de 2003 et 2009) révisé

Les textes enrichis d'attributs d'identification des items lexicaux et des ponctuations ont ensuite été convertis en XML et intégrés au corpus, où chacun est précédé d'un en-tête (cf. figure 4 : chaque élément <TEI> contient un élément <teiHeader> puis un élément <text>). La structuration des textes du corpus n'a pas été poussée aussi loin que le permet la *TEI* : les phrases sont délimitées, ainsi que les titres des textes et les notes lexicales et références bibliographiques fournies à la fin de certains d'entre eux, mais les dialogues par exemple ne le sont pas, dans la mesure où leur repérage ne semble indispensable ni à l'appariement des items du corpus et de ceux qui figurent en adresses dans le dictionnaire, ni à celui du sens attesté et de l'un de ceux qui sont décrits dans les subdivisions d'articles.

```
<TEI>
  <teiHeader><fileDesc><titleStmt><title level="a">2003p008+2009p008</title></titleStmt>
    <publicationStmt><availability status="restricted"><p>Usage limité, voir N. Gasiglia</p></availability></publicationStmt>
    <sourceDesc><bibl>première édition <date>2003</date> page <biblScope type="pp">008</biblScope>&#09;; seconde
      édition <date>2009</date> page <biblScope type="pp">008</biblScope></bibl></sourceDesc></fileDesc></teiHeader>
  <text><body>
    <head><s n="2003p008ph01">
<w n="2003p008ph1-1" lemma="mon" ana="DETPOSS">Ma</w>
<w n="2003p008ph1-2" lemma="sœur" ana="NCFS">sœur</w>
<w n="2003p008ph1-3" lemma="être" ana="VINDP3S">est</w>
<w n="2003p008ph1-4" lemma="un" ana="DETIFS">une</w>
<w n="2003p008ph1-5" lemma="sorcier" ana="NCFS">sorcière</w>
<c n="2003p008ph1-6" type="punctuation" ana="PCTFORTE">!</c></s></head>
<s n="2003p008ph02">
<w n="2003p008ph2-1" lemma="mon" ana="DETPOSS">Ma</w>
<w n="2003p008ph2-2" lemma="sœur" ana="NCFS">sœur</w>
<w n="2003p008ph2-3" lemma="être" ana="VINDP3S">est</w>
<w n="2003p008ph2-4" lemma="un" ana="DETIFS">une</w>
<w n="2003p008ph2-5" lemma="sorcier" ana="NCFS">sorcière</w>
<c n="2003p008ph2-6" type="punctuation" ana="PCTFORTE">.</c></s>
<s n="2003p008ph03">
<w n="2003p008ph3-1" lemma="on" ana="PPER3S">On</w>
<w n="2003p008ph3-2" lemma="ne" ana="ADV">ne</w>
<w n="2003p008ph3-3" lemma="savoir" ana="VINDP3S">sait</w>
<w n="2003p008ph3-4" lemma="pas" ana="ADV">pas</w>
<w n="2003p008ph3-5" lemma="comment" ana="SUB">comment</w>
<w n="2003p008ph3-6" lemma="ce" ana="PDS">c'</w>
<w n="2003p008ph3-7" lemma="être" ana="VINDP3S">est</w>
<w n="2003p008ph3-8" lemma="arriver" ana="VPARPMS">arrivé</w>
<c n="2003p008ph3-9" type="punctuation" ana="PCTFORTE">.</c></s>
```

Figure 4 : Produit de Cordial Analyseur (relatif au texte de la page 8 du manuel À portée de mots CE1 de 2003 et 2009) révisé et converti en XML

Intégrant les textes “longs” des éditions 2003 et 2009 du manuel *À portée de mots* CE1, le corpus compte 71 textes différents (éléments <text>), 1 963 phrases (<s>), 19 212 mots (<w>) et 3 784 ponctuations (<c>).

4 APPARIEMENT DES ITEMS LEXICAUX DU CORPUS ET DES ITEMS EN ADRESSES DU DICTIONNAIRE

La seconde phase d’annotation consiste à mettre en correspondance chaque item lexical de chaque texte avec un item utilisé comme adresse d’article dans le dictionnaire pour le cycle 2, s’il en existe qui soient valides. Cette tâche implique de relever la nomenclature du répertoire puis de rapprocher (i) les lemmes des items du corpus avec les formes des adresses et (ii) leurs codes morphosyntaxiques avec les catégorisations du dictionnaire. Au terme de ces appariements, les indications d’adressage introduites dans le balisage du corpus permettent d’évaluer la proportion de mots simples figurant dans celui-ci auxquels correspondent une adresse et donc une description dans le dictionnaire et, parmi eux, de ceux qui s’intègrent dans l’une des unités polylexicales qu’il décrit. Un retour critique sur les résultats obtenus permet de mettre en perspective les choix d’étiquetage et les résultats de la procédure semi-automatique de repérage des adresses d’articles pertinentes.

Les appariements des items du corpus et de ceux qui figurent en adresses du *Dictionnaire Hachette benjamin* ont le plus souvent pu être réalisés en se fiant seulement à la comparaison des formes graphiques des lemmes. Mais, comme le montre la figure 5, diverses mises en relation n’ont pas été aussi simples à établir :

Lemmes du corpus	Fréquence en corpus	Adresses DHB	Nombres de subdivisions	Adresses subordonnées
afin de	1	**afin**	[-]	
africain	1			
[…]		[…]		
Agathe	2			
[…]		[…]		
agenouiller	1	**agenouiller** (s')	[-]	
[…]		[…]		
agiter	4	**agiter**	[1.]	**s'agiter** [2.]
[…]		[…]		
Allemagne	1			
aller	82	**aller** (n.)	[-]	
		aller (v.)	[1.] [2.] [3.] [4.]	
[…]		[…]		
alors	30	**alors**	[1.]	**alors que** [2.]
[…]		[…]		
chou	8	**chou**	[1.]	**chou à la crème** [2.]
[…]		[…]		
chouette	1	**chouette** (n.)	[-]	
[…]		[…]		
être	489	**être** (n.)	[-]	
		être (v.)	[1.] [2.]	
[…]		[…]		
sorcier	20	**sorcier** (n.)	[-]	
		sorcière	[-]	

Figure 5 : Appariement des items à la nomenclature du Dictionnaire Hachette benjamin et des lemmes établis par Cordial Analyseur

1) certaines s'avèrent infaisables, parce que le dictionnaire n'a pas intégré de noms propres (prénoms, noms de pays, etc., comme *Agathe* ou *Allemagne*) à sa nomenclature, pas plus qu'il n'y accueille les déterminants, les pronoms, les nombres, les noms d'habitants de pays ou de continents ou les adjectifs (comme *africain*) qui expriment la relation existant entre le nom qualifié et un pays ou un continent (absents de la nomenclature, les nombres et les noms et adjectifs dérivés de noms propres géographiques sont toutefois présentés dans des paratextes : les premiers dans chaque édition avec un contenu identique – page 358 en 1996 et 2002, pages 378-379 en 2004 et 2007 –, les seconds seulement dans l'édition de 2002 – pages 558-559) ;

2) certaines impliquent la prise en compte de la catégorisation, en raison du fait que, pour certaines formes graphiques qui correspondent à des items de catégories différentes, le dictionnaire peut avoir choisi soit de ne retenir qu'un sous-ensemble de ceux-ci (pour *chouette*, l'adjectif, typique du parler enfantin et attesté dans le corpus – texte de 2003, page 38, phrase 24, items 7 à 17 : *ça avait l'air drôlement chouette d'être chercheur de choses*–, n'est pas décrit), soit de tous les prendre en compte (pour *aller* ou *être*, noms et verbes sont traités), ce qui empêche de procéder à l'appariement de chaque occurrence en corpus avec une adresse du dictionnaire sans contrôle de la correspondance du code morphosyntaxique porté par l'occurrence et de l'information catégorielle associée à l'item en adresse ;

3) certaines impliquent la confrontation des adresses subordonnées et du contexte des occurrences, parce que certains items sont décrits dans le dictionnaire comme pouvant entrer dans des unités polylexicales, comme *chou*, inclus dans *chou à la crème* décrit *s.v.* **chou** 2, et pour lequel le script d'appariement doit repérer la présence du nom simple mais aussi, le cas échéant, son insertion dans l'unité complexe (attestée dans le texte de 2003, page 86, phrase 12, items 27 à 30, cf. figure 6) ;

```
<w n="2003p086ph12-9" lemma="le" ana="DETDFS">La</w>
<w n="2003p086ph12-10" lemma="boulanger" ana="NCFS">boulangère</w>
[…]
<w n="2003p086ph12-16" lemma="avoir" ana="VINDI3S">avait</w>
[…]
<w n="2003p086ph12-23" lemma="le" ana="DETDMS">le</w>
<w n="2003p086ph12-24" lemma="cœur" ana="NCMS">cœur</w>
<w n="2003p086ph12-25" lemma="comme" ana="SUB">comme</w>
<w n="2003p086ph12-26" lemma="son" ana="DETPOSS">ses</w>
<w n="2003p086ph12-27" lemma="chou" ana="NCMP">choux</w>
<w n="2003p086ph12-28" lemma="à" ana="PREP">à</w>
<w n="2003p086ph12-29" lemma="le" ana="DETDFS">la</w>
<w n="2003p086ph12-30" lemma="crème" ana="NCSIG">crème</w>
<c n="2003p086ph12-31" type="punctuation" ana="PCTFORTE">.</c></s>
```

Figure 6 : Une occurrence du nom chou employé dans l'unité polylexicale chou à la crème

4) certaines impliquent la prise en compte du contexte immédiat, parce que, dans le dictionnaire, les verbes employés pronominalement sont décrits dans des articles autonomes quand ils n'ont qu'un emploi pronominal ou que celui-ci est le seul décrit mais dans une subdivision d'article introduite par une adresse subordonnée quand d'autres emplois sont décrits, alors que dans le corpus les formes verbales sont toutes lemmatisées de manière autonome, le pronom figurant dans l'élément <w> précédent en cas d'emploi pronominal (cf. figure 7) : si, par exemple, la mise en correspondance du lemme *agiter* (édition 2003, page 88, phrase 14, item 30) employé pronominalement (le lemme de l'item 29 est *se*) et de l'adresse homographe est facilitée par le fait que ce n'est que la seconde subdivision de l'article qui décrit l'emploi pronominal à apparier et que cet appariement sera réalisé de la même manière que celui de *chou à la crème* au point précédent, il n'en va pas de même pour celle du lemme *agenouiller* (édition 2009, page 78, phrase 15, item 12), alors même qu'il est employé pronominalement (le lemme de l'item 11 est *se*), avec l'adresse **s'agenouiller** du *Dictionnaire Hachette benjamin* (notée « s'**agenouiller** » dans les éditions de 1996 et 2002 et « **agenouiller** (s') », comme en figure 5, dans les éditions de 2004 et 2007) ;

```
<w n="2003p088ph14-22" lemma="le" ana="DETDMS">le</w>
<w n="2003p088ph14-23" lemma="liquide" ana="NCMS">liquide</w>
<w n="2003p088ph14-24" lemma="contenu" ana="ADJMS">contenu</w>
<w n="2003p088ph14-25" lemma="dans" ana="PREP">dans</w>
<w n="2003p088ph14-26" lemma="le" ana="DETDFS">l'</w>
<w n="2003p088ph14-27" lemma="oreille" ana="NCFS">oreille</w>
<w n="2003p088ph14-28" lemma="interne" ana="ADJSIG">interne</w>
<w n="2003p088ph14-29" lemma="se" ana="PPER3S">s'</w>
<w n="2003p088ph14-30" lemma="agiter" ana="VINDP3S">agite</w>
<w n="2003p088ph14-31" lemma="encore" ana="ADV">encore</w>
<w n="2003p088ph14-32" lemma="après" ana="PREP">après</w>
<w n="2003p088ph14-33" lemma="que" ana="SUB">que</w>
<w n="2003p088ph14-34" lemma="tu" ana="PPER2S">tu</w>
<w n="2003p088ph14-35" lemma="avoir" ana="VSUBP2S">aies</w>
<w n="2003p088ph14-36" lemma="finir" ana="VPARPMS">fini</w>
<w n="2003p088ph14-37" lemma="de" ana="PREP">de</w>
<w n="2003p088ph14-38" lemma="tourner" ana="VINF">tourner</w>
```

```
<w n="2009p078ph15-6" lemma="le" ana="DETDFS">l'</w>
<w n="2009p078ph15-7" lemma="infirmier" ana="NCFS">infirmière</w>
<w n="2009p078ph15-8" lemma="qui" ana="PRI">qui</w>
<w n="2009p078ph15-9" lemma="être" ana="VINDP3S">est</w>
<w n="2009p078ph15-10" lemma="venir" ana="VPARPFS">venue</w>
<w n="2009p078ph15-11" lemma="se" ana="PPER3S">s'</w>
<w n="2009p078ph15-12" lemma="agenouiller" ana="VINF">agenouiller</w>
<w n="2009p078ph15-13" lemma="à" ana="PREP">à</w>
<w n="2009p078ph15-14" lemma="mon" ana="DETPOSS">mes</w>
<w n="2009p078ph15-15" lemma="côté" ana="NCMP">côtés</w>
```

Figure 7 : Deux occurrences de verbes employés pronominalement dans le corpus

5) certaines impliquent de composer avec la diversité des traitements dictionnairiques ou avec celle des modes d'étiquetages, parce que certains items sont décrits dans le dictionnaire comme pouvant entrer dans des unités polylexicales, comme *alors* inclus dans *alors que* décrit *s.v.* **alors** 2, tandis que d'autres ne le sont pas, comme *afin* décrit dans un article à une seule subdivision de description dont la contextualisation présente une occurrence de *afin que* mise en relief en gras mais ni traitée en adresse ni spécifiquement décrite, et qu'il faut mettre en cohérence les modes de traitement du dictionnaire et les découpages accomplis par Cordial Analyseur, qui a induit que, dans le corpus (cf. figure 8), les items *alors* et *que* sont isolés, contenus dans deux éléments <w> successifs (texte de 2003, page 78, phrase 8, items 15 et 16), mais que *afin de* constitue le contenu d'un seul élément (texte de 2003, page 84, phrase 20, item 19) ;

```
<w n="2003p078ph8-7" lemma="Margot" ana="NPFS">Margot</w>
<w n="2003p078ph8-8" lemma="ne" ana="ADV">n'</w>
<w n="2003p078ph8-9" lemma="avoir" ana="VINDP3S">a</w>
<w n="2003p078ph8-10" lemma="même" ana="ADV">même</w>
<w n="2003p078ph8-11" lemma="pas" ana="ADV">pas</w>
<w n="2003p078ph8-12" lemma="monter" ana="VPARPMS">monté</w>
<w n="2003p078ph8-13" lemma="un" ana="DETIFS">une</w>
<w n="2003p078ph8-14" lemma="marche" ana="NCFS">marche</w>
<w n="2003p078ph8-15" lemma="alors" ana="ADV">alors</w>
<w n="2003p078ph8-16" lemma="que" ana="PRI">que</w>
<w n="2003p078ph8-17" lemma="le" ana="DETDPIG">les</w>
<w n="2003p078ph8-18" lemma="autre" ana="PIPIG">autres</w>
<w n="2003p078ph8-19" lemma="être" ana="VINDP3P">sont</w>
<w n="2003p078ph8-20" lemma="déjà" ana="ADV">déjà</w>
<w n="2003p078ph8-21" lemma="presque" ana="ADV">presque</w>
<w n="2003p078ph8-22" lemma="arrivé" ana="ADJMP">arrivés</w>
<w n="2003p078ph8-23" lemma="au" ana="DETDMS">au</w>
<w n="2003p078ph8-24" lemma="premier" ana="ADJORD">premier</w>
<w n="2003p078ph8-25" lemma="étage" ana="NCMS">étage</w>
```

```
<w n="2003p084ph20-2" lemma="le" ana="DETDPIG">Les</w>
<w n="2003p084ph20-3" lemma="zèbre" ana="NCMP">zèbres</w>
<w n="2003p084ph20-4" lemma="kilongais" ana="ADJI">kilongais</w>
<w n="2003p084ph20-5" lemma="être" ana="VINDP3P">sont</w>
<w n="2003p084ph20-6" lemma="vert" ana="ADJMP">verts</w>
<w n="2003p084ph20-7" lemma="à" ana="PREP">à</w>
<w n="2003p084ph20-8" lemma="rayure" ana="NCFP">rayures</w>
<w n="2003p084ph20-9" lemma="blanc" ana="ADJFP">blanches</w>
[…]
<w n="2003p084ph20-19" lemma="afin de" ana="PREP">afin de</w>
<w n="2003p084ph20-20" lemma="servir" ana="VINF">servir</w>
<w n="2003p084ph20-21" lemma="de" ana="PREP">de</w>
<w n="2003p084ph20-22" lemma="passage" ana="NCMS">passage</w>
<w n="2003p084ph20-23" lemma="pour" ana="PREP">pour</w>
<w n="2003p084ph20-24" lemma="piéton" ana="ADJMP">piétons</w>
<w n="2003p084ph20-25" lemma="quand" ana="SUB">quand</w>
<w n="2003p084ph20-26" lemma="il" ana="PPER3P">ils</w>
<w n="2003p084ph20-27" lemma="se" ana="PPER3S">se</w>
<w n="2003p084ph20-28" lemma="coucher" ana="VINDP3P">couchent</w>
```

Figure 8 : Deux occurrences de locutions balisées différemment

6) enfin, certains problèmes découlent du paramétrage du lemmatiseur, qui ne tient pas compte des usages lexicographiques : ils concernent par exemple le nom *sœur* lemmatisé par *frère* et corrigé manuellement (cf. § 3.) puisque les dictionnaires s'accordent à considérer qu'il s'agit de deux unités linguistiques distinctes qui méritent d'être traitées individuellement (cf. figure 3, phrases 1 et 2, items 2), mais aussi d'autres noms féminins qui ont été lemmatisés au masculin (comme *sorcière* en figure 3, phrases 1 et 2, items 5) sans que nous ayons considéré que ce soit problématique alors que le *Dictionnaire Hachette benjamin* (comme le *Larousse des débutants* mais pas le *Robert benjamin* ni les répertoires pour le cycle 3 des trois éditeurs) propose un article pour le nom masculin et un autre pour le nom féminin avec des descriptions de sens différentes (cf. figure 9), ce qui invalidera les futures mises en relation du mot en contexte avec l'article qui décrit son sens (en l'occurrence, effectuer une correction aurait été trop indexé sur le mode de traitement d'un sous-ensemble minoritaire de dictionnaires pour être pertinent).

un **sorcier** (NOM)
Personne qui pratique la magie.
Dans certains pays, les gens croient que les ***sorciers*** *peuvent faire des choses extraordinaires.*

une **sorcière** (NOM)
Dans les contes,
femme souvent méchante et laide,
et qui a des pouvoirs magiques.
Myriam a lu l'histoire d'une vieille ***sorcière*** *qui changeait les enfants en crapauds.*
➡ Regarde page 90 : **à califourchon**.

Figure 9 : Articles ***sorcier*** *et* ***sorcière*** *du Dictionnaire Hachette benjamin de 1996 (depuis la refonte iconographique de 2004, le renvoi final s.v.* ***sorcière*** *a disparu)*

Une fois réalisés les appariements des lemmes des items du corpus et des adresses du dictionnaire, ces derniers viennent enrichir le balisage des éléments <w> du corpus dont ils constituent les valeurs de l'attribut @lemmaRef (cf. figure 10).

```
<head><s n="2003p008ph01">
   <w n="2003p008ph1-1" lemma="mon" ana="DETPOSS">Ma</w>
   <w n="2003p008ph1-2" lemma="sœur" ana="NCFS" lemmaRef="DHB96/sœur">sœur</w>
   <w n="2003p008ph1-3" lemma="être" ana="VINDP3S" lemmaRef="DHB96/être_V">est</w>
   <w n="2003p008ph1-4" lemma="un" ana="DETIFS">une</w>
   <w n="2003p008ph1-5" lemma="sorcier" ana="NCFS" lemmaRef="DHB96/sorcier">sorcière</w>
   <c n="2003p008ph1-6" type="punctuation" ana="PCTFORTE">!</c></s></head>
<s n="2003p008ph02">
   <w n="2003p008ph2-1" lemma="mon" ana="DETPOSS">Ma</w>
   <w n="2003p008ph2-2" lemma="sœur" ana="NCFS" lemmaRef="DHB96/sœur">sœur</w>
   <w n="2003p008ph2-3" lemma="être" ana="VINDP3S" lemmaRef="DHB96/être_V">est</w>
   <w n="2003p008ph2-4" lemma="un" ana="DETIFS">une</w>
   <w n="2003p008ph2-5" lemma="sorcier" ana="NCFS" lemmaRef="DHB96/sorcier">sorcière</w>
   <c n="2003p008ph2-6" type="punctuation" ana="PCTFORTE">.</c></s>
<s n="2003p008ph03">
   <w n="2003p008ph3-1" lemma="on" ana="PPER3S" lemmaRef="DHB96/on_MINV">On</w>
   <w n="2003p008ph3-2" lemma="ne" ana="ADV">ne</w>
   <w n="2003p008ph3-3" lemma="savoir" ana="VINDP3S" lemmaRef="DHB96/savoir">sait</w>
   <w n="2003p008ph3-4" lemma="pas" ana="ADV" lemmaRef="DHB96/pas_MINV">pas</w>
   <w n="2003p008ph3-5" lemma="comment" ana="SUB" lemmaRef="DHB96/comment">comment</w>
   <w n="2003p008ph3-6" lemma="ce" ana="PDS">c'</w>
   <w n="2003p008ph3-7" lemma="être" ana="VINDP3S" lemmaRef="DHB96/être_V">est</w>
   <w n="2003p008ph3-8" lemma="arriver" ana="VPARPMS" lemmaRef="DHB96/arriver">arrivé</w>
   <c n="2003p008ph3-9" type="punctuation" ana="PCTFORTE">.</c></s>
```

Figure 10 : Texte de la figure 4 avec appariement des lemmes des éléments <w> et des items en adresse dans le Dictionnaire Hachette benjamin de 1996

5 APPARIEMENT DES ITEMS LEXICAUX DU CORPUS ET DES ORDONNATEURS DE DESCRIPTIONS DE SENS DES ARTICLES QUI DÉCRIVENT DES MOTS POLYSÉMIQUES

La dernière phase d'annotation consiste, pour chaque item présent dans le corpus et correspondant à une adresse dans le dictionnaire, à chercher si la compréhension de chaque occurrence de l'item dans son contexte initial est éclairée par la ou l'une des description(s) du dictionnaire (cf. figure 11, où, par exemple, le verbe *être* employé comme troisième item dans les phrases 1 et 2 y a une acception décrite dans la première division de l'article, alors que son emploi comme auxiliaire en phrase 3 item 7 n'est pas envisagé dans celui-ci). Cette mise en relation des items en contexte avec les descriptions pertinentes, en supposant qu'il en existe bien, permettra vraisemblablement d'établir certains des critères déterminant les choix opérés manuellement qui sont opératoires pour le développement de la fonction de sélection de description qui pourrait intégrer le module de traitement des hyperappels.

Cette partie du traitement n'a été qu'ébauchée à ce stade de l'élaboration du corpus. C'est la plus longue et la plus minutieuse des tâches à effectuer si nous voulons établir quels sont les indices de chaque contexte et de chaque description de sens qui permettent d'établir les appariements pertinents et à terme de déterminer les modalités de mise en œuvre d'un module d'hyperappel du dictionnaire à partir des textes de ce corpus mais aussi, ultérieurement, d'autres textes automatiquement annotés dès lors qu'ils sont affichés à l'écran pour être lus.

```
<head><s n="2003p008ph01">
   <w n="2003p008ph1-1" lemma="mon" ana="DETPOSS">Ma</w>
   <w n="2003p008ph1-2" lemma="sœur" ana="NCFS" lemmaRef="DHB96/sœur">sœur</w>
   <w n="2003p008ph1-3" lemma="être" ana="VINDP3S" lemmaRef="DHB96/être_V#1">est</w>
   <w n="2003p008ph1-4" lemma="un" ana="DETIFS">une</w>
   <w n="2003p008ph1-5" lemma="sorcier" ana="NCFS" lemmaRef="DHB96/sorcier">sorcière</w>
   <c n="2003p008ph1-6" type="punctuation" ana="PCTFORTE">!</c></s></head>
<s n="2003p008ph02">
   <w n="2003p008ph2-1" lemma="mon" ana="DETPOSS">Ma</w>
   <w n="2003p008ph2-2" lemma="sœur" ana="NCFS" lemmaRef="DHB96/sœur">sœur</w>
   <w n="2003p008ph2-3" lemma="être" ana="VINDP3S" lemmaRef="DHB96/être_V#1">est</w>
   <w n="2003p008ph2-4" lemma="un" ana="DETIFS">une</w>
   <w n="2003p008ph2-5" lemma="sorcier" ana="NCFS" lemmaRef="DHB96/sorcier">sorcière</w>
   <c n="2003p008ph2-6" type="punctuation" ana="PCTFORTE">.</c></s>
<s n="2003p008ph03">
   <w n="2003p008ph3-1" lemma="on" ana="PPER3S" lemmaRef="DHB96/on_MINV">On</w>
   <w n="2003p008ph3-2" lemma="ne" ana="ADV">ne</w>
   <w n="2003p008ph3-3" lemma="savoir" ana="VINDP3S" lemmaRef="DHB96/savoir#2">sait</w>
   <w n="2003p008ph3-4" lemma="pas" ana="ADV" lemmaRef="DHB96/pas_MINV">pas</w>
   <w n="2003p008ph3-5" lemma="comment" ana="SUB" lemmaRef="DHB96/comment#1">comment</w>
   <w n="2003p008ph3-6" lemma="ce" ana="PDS">c'</w>
   <w n="2003p008ph3-7" lemma="être" ana="VINDP3S" lemmaRef="DHB96/être_V#ABS">est</w>
   <w n="2003p008ph3-8" lemma="arriver" ana="VPARPMS" lemmaRef="DHB96/arriver#2">arrivé</w>
   <c n="2003p008ph3-9" type="punctuation" ana="PCTFORTE">.</c></s>
```

Figure 11 : Texte de la figure 10 avec indication des ordonnateurs de descriptions pertinentes des items en adresse dans le Dictionnaire Hachette benjamin de 1996

6 PERSPECTIVES

Outre la documentation de l'étude métalexicographique d'un dictionnaire scolaire imprimé récent (le *Dictionnaire Hachette benjamin* a été publié en 1996 et réédité en 2002, 2004 et 2007), centrée sur les services qu'il peut effectivement rendre aux élèves qui savent le manipuler pour y chercher des indications susceptibles d'éclairer leur compréhension de certains items des textes lus, le corpus élaboré doit fournir des éléments de réflexion en vue de la mise en place d'un hyperappel de dictionnaire électronique qui aiderait véritablement les élèves les plus en difficulté en leur permettant (i) d'accéder à une description du mot problématique sans avoir à spécifier eux-mêmes quels sont le lemme et la catégorie grammaticale de cet item (ce qui peut être particulièrement malaisé quand la forme est ambiguë et qu'elle n'est pas interprétée) et (ii) d'être directement orientés, au sein de la description du mot cherché, vers les indications qui leur seront utiles. Les données obtenues à l'issue de la seconde phase d'annotation (cf. § 4.) doivent renseigner le premier point en fournissant des matériaux pour réfléchir à l'alignement des étiquetages produits par des outils comme Cordial Analyseur et des adresses de dictionnaires. L'exploitation des annotations de la troisième phase (cf. § 5.) consiste à établir ce qui permet de déterminer si une description est pertinente, ou de choisir celle qui l'est si plusieurs sont disponibles, afin de documenter la conception d'une procédure de sélection automatique des descriptions à partir des contextes motivant la consultation du dictionnaire, ce qui pourrait en outre conduire à réfléchir à ce qui mériterait d'être explicitement présenté dans les articles afin d'aider les élèves à évaluer si la sélection de description établie par le module d'hyperappel est valide et, dans la négative, à trouver les repères utiles pour la sélection d'une meilleure description (cf. Gasiglia (soumis)). Cette étude nécessitera certainement de prendre en compte les modes de description de différents dictionnaires, de manière à observer si certains s'avèrent plus efficaces, ce qui conduira à répéter les appariements précédents pour d'autres textes dictionnairiques et ainsi à enrichir notre connaissance des dictionnaires scolaires actuels et à stimuler les réflexions relatives à ce que pourraient être les dictionnaires électroniques des futurs environnements numériques de travail des élèves du primaire.

7 RÉFÉRENCES

Dictionnaires et manuels utilisés

À portée de mots pour le CE1 = Leclec'h-Lucas J., Lucas J.-C. et Meunier R. (2003). *Français CE1 cycle 2 niveau 3. À portée de mots*, Paris : Hachette Éducation ; nouv. éd., 2009.

Dictionnaire Hachette benjamin, Paris : Hachette Éducation, 1996, 2002 ; nouv. éd., 2004, 2007.

Dictionnaire Hachette junior, Paris : Hachette Éducation, 2006.

Larousse des débutants, Paris : Larousse, 2005.

Larousse junior, Paris : Larousse, 2008.

Le Robert benjamin, Paris : Dictionnaires Le Robert, 2009.

Le Robert junior illustré, Paris : Dictionnaires Le Robert, 2005.

Autres références

Corbin P. et Gasiglia N. (à paraître). « Éléments pour un état de la description de la variété des usages lexicaux dans les dictionnaires français monolingues actuels ». Actes du colloque international *La marque lexicographique : quel avenir ?*, Université de Chypre, 21-23 octobre 2006.

Gasiglia N. (2008a). « Le traitement des emprunts dans les dictionnaires d'apprentissage français : options descriptives et choix rédactionnels ». Dans J. Pruvost (dir.), *Les Journées des dictionnaires de Cergy. Dictionnaires et mots voyageurs. Les 40 ans du Petit Robert. De Paul Robert à Alain Rey,* Actes de la Journée des dictionnaires du 14 mars 2007, collection Actes de colloque, Éragny-sur-Oise : Éditions des Silves. p. 153-212.

Gasiglia N. (2008b). « Le traitement des anglicismes dans quelques dictionnaires français pour jeunes lecteurs ». Dans F. Maniez et P. Dury (dir.), *Lexicographie et terminologie : histoire de mots. Hommage à Henri Béjoint*, Travaux du CRTT, Gap : Louis Jean imprimeur. p. 57-174.

Gasiglia N. (2008c). « Description of loan words in French school dictionaries: treatment of words of foreign origin in *Dictionnaire Hachette junior* (2006) and *Le Robert junior illustré* (2005) ». Dans E. Bernal et J. DeCesaris (éds.), *Proceedings of the XIII EURALEX International Congress (Barcelona 15-19 July 2008)*, Sèrie Activitats 20, Barcelona : Institut Universitari de Lingüística Aplicada. p. 1115-1122.

Gasiglia N. (2008d). « Le traitement des "identifications diachroniques" dans des dictionnaires scolaires français : évaluation de pratiques et proposition de principes de rédaction alternatifs ». Dans J. Durand, B. Habert et B. Laks (resp.), *Congrès mondial de linguistique française. Paris, 9-12 juillet 2008*, Paris : Institut de Linguistique Française / EDP Sciences. p. 1117-1136 du CD-Rom ; éd. en ligne : http://www.linguistiquefrancaise.org/articles/cmlf/pdf/2008/01/cmlf08236.pdf. p. 1129-1148.

Gasiglia N. (2009). « Évolutions informatiques en lexicographie : ce qui a changé et ce qui pourrait émerger ». *Lexique* 19, p. 224-298.

Gasiglia N. (à paraître a). « Some editorial orientations for a multi-tier electronic monolingual school dictionary ». Dans S. Granger et M. Paquot (éds.), Actes colloque *eLexicography in the 21st century: New challenges, new applications*, Université de Louvain-la-Neuve (Belgique), 22-24 octobre 2009.

Gasiglia N. (à paraître b). « Étude des marquages d'emplois lexicaux identifiés comme relevant du "parler enfantin" dans des dictionnaires monolingues français contemporains ». Dans C. Rey et Ph. Reynes (dir.), Actes du colloque Atelier du LESCLaP *Dimensions (visions et représentations) sociolinguistiques dans les dictionnaires*, Université de Picardie Jules Verne, Amiens, 9 décembre 2009.

Gasiglia N. (à paraître c). « Les divers aspects de la prescription dans les dictionnaires scolaires ». Dans D. Candel et D. Kibbee (éds.), Actes du colloque international *Prescriptions en langue*,

Paris, 15-16 novembre 2007.
Gasiglia N. (soumis). « Le *Dictionnaire Hachette benjamin* met-il à la portée des élèves du CE1 les mots des textes d'un manuel de français ? ». *Lexique*.
TEI: P5 Guidelines, http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

THE FREQUENCY OF WRITTEN AND SPOKEN ANGLICISMS IN TWO VARIETIES OF FRENCH

Jesse Harris and Walcir Cardoso
Concordia University & The Centre for the Study of Learning and Performance

SUMMARY

The goal of this study is to examine the frequency of anglicisms in both written and spoken French using a corpus collected from two reality television shows and from Internet blogs - data representing two varieties of French: Quebec French (QF), and France French (FF), in order to determine which variety contains a higher frequency of anglicisms. The following research questions guided this study: (1) Which variety of French uses a higher percentage of anglicisms?; (2) Will one language mode (written vs. spoken French) yield a higher frequency of anglicisms than the other?; and (3) What is the effect of a type frequency versus a token frequency analysis for the quantification of anglicism in the corpora analyzed?

The corpus (including oral and written data) of more than 40,000 words was designed especially for this study in order to control for certain variables. Data for spoken French for each variety were gathered from two reality television programs: Star Academy (2008 season in France) and Star Académie (2009 season in Quebec), two versions of the same program (same concept and format) with the same target audience (for the sake of comparability). In order to acquire writing samples from French speakers, current text from various weblogs (blogs) was collected.

Based on the analysis of the data, it appears that anglicisms represent 0.99% of the total corpus (including both FF and QF). In other words, the results of our analysis indicate that in 2008/9, anglicisms tend to occur less than one percent (1%) of the time. Furthermore, of this total, the percentage of anglicism tokens in FF was 0.75% of the entire corpus, while tokens in QF totaled 1.23%, suggesting that in this study, QF contains more anglicisms than FF. However, focusing on language modes, the results revealed an interesting pattern in which FF had the highest incidence of anglicisms in the written corpus (blogs), while QF favored anglicisms in the oral corpus (TV).

In sum, the overall small frequency of anglicisms in the French language may provide an answer to language purists regarding the level of "infiltration" of the English language into French. Comparing with “common knowledge” and decades-old corpus analyses (e.g., Forgue, 1986; Rey-Debov, 1980), the frequency of anglicism in contemporary French is comparable to decades old studies on anglicism frequency (i.e., Forgue, 1986; Mareschal, 1992; and Théoret, 1991), and remains considerably lower today than is commonly thought.

1 INTRODUCTION

Over the decades, linguists around the world have studied the infiltration and usage of one language in another language. Some view this type of cross-linguistic influence as interference (see Mougeon & Beniak, 1991; Weinreich, 1974), while others see borrowings as a source of enrichment to the language (see Guiraud, 1965; Le Prat, 1980; Picone, 1996). Regardless of the various sentiments towards the phenomenon, given the long trail of

borrowings throughout linguistic history, it is safe to assume that languages will continue to borrow from each other well into the future.

The borrowing of English words in particular has interested many scholars, especially speakers of French. Historically speaking, Wise (1997) points out that the borrowing of English words into French dates back to the 1600s with a significant increase in anglicisms starting by the end of the 18th century. De Ullmann (1947) even cites a very small number of English borrowings before the 1600s with one instance (alderman) first appearing in French in 1363.

Eventually, English borrowings began to create worry amongst certain French-speaking communities ultimately resulting in legal measures taken to protect the language from English "contamination". According to Nadeau and Barlow (2006), Quebec spearheaded language protection after World War II in 1959, which eventually lead to the Loi 101: Charter of the French Language in 1976. In fact, it was France that then modeled their well-known language protection policies after the movement in Quebec. Previous language protection attempts in France had proved largely ineffective up to that point due to the lack of belief by French authorities that there truly was a threat posed by the English language. However, French literary critic René Étimble's (1964) Parlez-vous franglais? opened many French speakers' eyes to the prevalence of English in France spawning the founding of the Délégation générale à la langue française et aux langues de France (DGLFLF) in the 1990's, a committee charged with the protection of the French language in France .

And so, has the effort put forth by the governments of Quebec and France to guard against outside English language elements influenced the actual everyday usage of English words and phrases in the French language today? What kind of role do words like "feeling" and "tripper" play in the informal, unmonitored speech of a typical Francophone? In order to answer these questions, one must first consider the actual amounts and percentages of anglicisms in the language as a whole. In this study I will seek to address the question of anglicism frequency and distribution in French. The next section will provide an overview of pertinent research and findings on the topic of anglicisms in French followed by a terminology section on anglicisms in general and relevant background. Next, three research questions will be presented followed by a detailed section on variable selection and corpus design. A description of data manipulation and analysis will follow, and finally the paper will conclude with a summary as well as a brief discussion on possible contributions of the proposed research.

2 PREVIOUS RESEARCH

In order to investigate more specifically the nature of borrowings across languages, this study focuses on borrowings between two languages in particular: English as the donor language and French as the recipient language. The political "hotness" of the anglicism topic in both Quebec and France generated a host of research and critiques from linguists in both regions particularly in the 1970's-1990's (Nadeau & Barlow, 2006). Book chapters and journal articles dedicated to the topic of anglicisms, however, have been largely descriptive, devoting energy primarily to explaining the nature of and reasons for English borrowings in French. Less frequent have been research dedicated to the pure numbers. The following section briefly reviews the empirical studies that have already addressed, from varied perspectives, the frequency of anglicisms in French.

The number of studies devoted exclusively to quantifying anglicisms in French is somewhat sparse. Authors Forgue (1986), Théoret (1991) and Mareschal (1992) all took

interest in this linguistic phenomenon, considering a number of common variables important to the investigation of anglicisms in French.

One important variable considered by each author was the variety of French to study. As different varieties of French are inevitably attached to different cultural, political, and historical events and norms, language variety not withstanding, the use and frequency of anglicisms could conceivably vary from region to region. In Forgue's (1986) study, for example, the author undertook the task of identifying anglicisms in the variety of French spoken in France. Théoret (1991) also investigated a single variety of French from various regions of the province of Quebec in Canada (Estrie, Montreal, Quebec, and Saguenay-Lac St-Jean). Finally, in Mareschal (1992), the author presented results from research on anglicisms in four different francophone regions: Belgium, France, Quebec, and Switzerland.

The number and the source of the words studied varied between authors as well. Forgue (1986) analyzed just under fifteen million words (1,370,000 words), and his corpus consisted of daily articles from the French newspaper Le Monde in 1977. These newspaper articles provided data for only one language mode: written language. The potentially problematic nature of this fact, as admitted by Forgue himself, is that Le Monde was a newspaper generally seen as "elitist" employing a highly formal register of language. The likelihood that the language in this newspaper (including anglicization) was a representative sample of current day language at the time of the study is slim. In Théoret (1991), the corpus was considerably smaller than that of Forgue (1986). The Sherbrooke Corpus was a corpus of one million words made up fifty percent of "spontaneous oral" language and fifty percent "non-spontaneous oral" language (i.e., language "written to be spoken": folklore, theatre, radio broadcasts, soap operas, monologues, etc.). Like Forgue, the Sherbrooke corpus represented only one mode of language, comprised only of spoken (not written) language. Although the author does not disclose the total size of the corpora, Mareschal (1992), as with Forgue (1986), also gathered data for her research from newspapers during the late 1970s and early 1980s. Unlike Forgue, the author used newspapers accessible to the general public.

On the other hand, all three authors employed a similar approach to quantifying the linguistic variable. This was a dual calculation approach which took into consideration anglicism word tokens as well as word types. For example, for the token calculation Théoret (1991) conducted a simple frequency count of the total number of individual anglicism occurrences versus the total number of word occurrences in the corpus. Additionally, the author derived an anglicism token count by tallying the total number of anglicism word types (n = 699) in the corpus and dividing by the total number of French word types (n = 11,327). Forgue (1986) and Mareschal (1992) also considered both anglicism tokens and types in their analyses. The obvious importance of taking both word tokens and word types into consideration in any frequency analysis will be discuss further in the forthcoming section "Variable Selection and Corpus Design".

Of course, the results and their implications proved the most interesting comparisons between these three authors. The results of Forgue's (1986) research produced a total of 8,200 anglicism tokens, translating into roughly 0.60% of the collected corpus. Moreover, when counting types (n = 680), this figure dropped to 0.04%. Théoret's (1991) frequency count unveiled 2,861 anglicism tokens in the corpus of one million French words (0.28%), however for the analysis of type of anglicism, the result was a figure of 6% indicating that six percent of the total distinct word types in the corpus (English or French) were comprised of anglicisms. The results of Mareschal's (1992) analysis uncovered 1,801 total tokens of anglicisms and 904 total distinct anglicism word types. Unique from Forgue and Théoret,

Mareschal compared the distribution percentages of anglicisms between French-speaking regions rather than measuring the total percentage of anglicisms in the French language.

The relevant information and results from the above literature review are summarized in Table 1.

Author	**Forgue** (1986)	**Théoret** (1991)	**Mareschal** (1992)
Corpus size	15 M	1 M	?
Source	Written (*Le Monde* daily, 1977)	Oral (Sherbrooke corpus)	Written (Daily newspapers)
Variety	France	Quebec	Belgium, France, Quebec, Switzerland
Results	0.6% / 0.04% (tokens/ types)	0.28% / 6.0% (tokens/types)	n = 1,801 / 904 (tokens/ types)

Table 1: Summary of previous anglicism research

The design and results of these studies are revealing in many ways regarding the variables adopted. However, to date, there still exists a gap in the literature on anglicism frequency in French. The most salient inadequacy is the timeliness of the data. Not only were the primary studies on anglicism frequency carried out over two decades ago, but also the data in these studies were already somewhat obsolete at the time of publication. Another shortcoming involves the inconsistency in variables studied across the literature. While some authors choose to look at only one mode (written or spoken language) across several regions (language varieties), other authors focus only on one language mode in one language variety.

We are not aware of any study to date that incorporates current data into research on both language modes across more than one language variety. Consequently, drawing from method and design present and/or lacking in the research reviewed here, the objective of the current study is to collect and analyze anglicism frequency data from up-to-date sources in both written and spoken language across two distinct varieties of French.

3 TERMINOLOGY: WHAT'S IN A NAME?

To date, there has been a substantial amount of research on the topic of anglicisms and their integration in various languages. General definitions for anglicisms have been embedded within literature and empirical studies, and even dictionary forwards and introductions provide explanations on the topic (see the *Petit Robert*, 2005; the *Dictionnaire de Français Plus*, 1988; Villers, 1988; and Rey-DeBove & Gagnon, 1980). Of the available literature, however, only a small number of books and articles on the topic have attempted a clear definition. What is more, of the works that actually define "anglicism", very few have gone beyond a general explanation to consider the depth or breadth of the variety of important linguistic factors contributing to the identification of an anglicism (with some exceptions: see Mareschal, 1992, and Picone, 1996 for comprehensive categorizations of anglicisms). As a result, confusion in the naming terminology (and presumably the identification) of various borrowings from English ensues.

3.1 Anglicism categories

An exploratory investigation and synthesis of the term "anglicism" and its various definitions in the literature has revealed approximately six different categories of anglicisms in the French language: *wholesale anglicisms*, *direct translations*, *semantic anglicisms*, *hybrids, French inventions and modifications,* and *morphological anglicisms*. The next section will

focus more specifically on the nuances and complexities of these six different kinds of anglicisms[9].

3.1.1 WHOLESALE ANGLICISMS

Wholesale anglicisms (or *Intact/quasi-intact borrowing*, *whole/partial borrowing*, *conscious borrowing*, *direct loan*, and *Frenchified anglicisms*), are a type of anglicism that undergoes (virtually) no change from English to French (see Bouchard, 1999; Forgue, 1986; Grigg, 1997; Guiraud, 1965; Mareschal, 1992; Rifelj, 1996; Spence, 1989; Trescases, 1982; Villers, 1988). The word or expression in English is identical in both form and meaning to its usage in French. For example the English word "weekend" is used with identical orthography and meaning in French. The same is true for "muffin", "fair play", and "bowling". That is to say that the French language imports an English object as well as its corresponding English meaning.

3.1.2 DIRECT TRANSLATIONS

A second type of anglicism is English words, expressions, or ideas directly translated from English (usually morpheme for morpheme) into French (see Forgue, 1986; Grigg, 1997; Guiraud, 1965; Mareschal, 1992; de Ullmann, 1947; Wise, 1997). These direct translations (most commonly called *calques*, *structural calques,* or *loan translations*) may include "gratte-ciel" for *sky scraper*, "haut parleur" for *loud speaker* or even "bienvenue" for *welcome*, the short form of *you're welcome*[10]. Grigg (1997) as well as Mareschal (1992) and Picone (1996) also look at the direct translation of English compounds into French. According to the literature, this syntactic process can happen in one of two ways: either the English compound is borrowed into French taking on the French word order as in "facteur-risque" (for *risk factor*), or the English compound is borrowed into French imitating the English compounding structure (with a left-headed modifier) such as "télérepas" (for *TV dinner*). Thus, direct translation is essentially the borrowing of an English object, its meaning, and occasionally its syntax while leaving behind the English name.

3.1.3 SEMANTIC ANGLICISMS

By and large, the most difficult type of anglicism to identify is the semantic anglicism (or *semantic borrowing*, *semantic calque*, *loan shift*, and *semantic imitation*). This type of anglicism constitutes borrowing an English word that is similar to an already existent French word (with a different meaning), and superimposing the English meaning on the French word (see Bouchard, 1999; Guiraud, 1965; Mareschal, 1992; Picone, 1996; Rifelj, 1996; de Ullmann, 1947; Villers, 1988; Wise, 1997). The most common (and many times the only) example of this type of anglicism in the literature is the French verb "réaliser" being incorrectly employed with the English meaning *to realize*. False cognates like these are the most susceptible kinds of words to this kind of anglicization. In French, the traditional usage of the verb "réaliser" means *to fulfill* or *to achieve*. However, when "réaliser" undergoes semantic anglicisation, it takes on the English definition/meaning of *to realize* (which already

[9] Although divided here into distinct and separate categories, it is often the case that anglicisms produced in both speech and writing overlap and prove to be composed of elements from more than one of these categories.

[10] An interesting stipulation brought up by Grigg (1997) is that oftentimes, one needs to have certain knowledge of current English culture in order to understand the directly translated French expression. Two examples are "chasseur de têtes" and "parfum du jour". Despite the fact that the expression uses entirely French words, a person would need to be familiar with the workforce recruitment agents known as *head-hunters*, to understand. Nor would one necessarily understand the concept of *flavor of the day* commonly used to refer to certain individuals who change love interests on a frequent basis.

exists as “se rendre compte de” in French)[11]. Because of the subtlety in meaning change as well as the natural difficulty caused by false cognates, semantic anglicisms prove difficult to detect and often go unnoticed even by native speakers of French.

3.1.4 HYBRIDS

A small category of anglicisms includes the hybrid as discussed by Grigg (1997), Picone (1996) and Trescases (1982). This anglicism type combines existing French elements with borrowed English words. For example, the hybrid “surbooker” is composed of the English verb *to book* as well as the French prefix “sur” meaning *over*. This results in an anglicism meaning to *overbook*. At first sight, hybrids resemble direct translations and wholesale anglicisms. However, whereas direct translations and wholesale anglicisms borrow all elements of the English word or expression into French, hybrids borrow an English word as a base and insert existing French elements around the base. Other examples of hybrids in the literature include “en live” and “top modèle”. Ultimately, one could say that a hybrid maintains the same meaning from English to French, and that the form is a mix between a direct translation and a wholesale anglicism.

3.1.5 FRENCH INVENTIONS AND MODIFICATIONS

The next type of anglicism lies at the opposite end of the spectrum from wholesale anglicisms in that there is no transfer of meaning from English to French. A French invention (also known as *false anglicisms*, *anglicisms of the signifier*, *pseudo borrowing*, *pseudo-anglicisms*, and *over anglicization*) is a word based on English elements that adopts a French meaning which is unusual or unknown to Anglophones (see Forgue, 1986; Grigg, 1997; Guiraud, 1965; Mareschal, 1992; Picone, 1996; Spence, 1989; Thogmartin, 1984; Trescases, 1982; Villers, 1988). English words are given a French meaning that is distorted or does not equate to the meaning of the word in English. For example, the word “tennisman” would appear quite familiar to a native English speaker due to the two English elements *tennis* and *man*. Though one may be able to guess the French meaning of this word, *tennis player* remains the conventional term in English. Another less intuitive example of a French invention is the popular word “footing”. Although the English word *foot* and the English morpheme “ing” are both evident, an anglophone may not necessarily know to bring sneakers when invited to go “footing” (*go for a jog*). Finally, a French speaker may talk about “un lifté” who lives down the street. The English word *lift* is perceptible, but in what sense of the word? In fact, “un lifté” refers to a person who has had a *face lift*; surgically speaking they have been *lifted*.

Another group of anglicisms explored mainly by Grigg (1997) involves the influence of French lexis and syntax on English borrowings. One instance of this type of modification occurs through the truncation (shortening) of an English word to make a French word (see also Mareschal, 1992). For example, a French person could very well put on a “sweat” (*sweatshirt*), grab their “walk” (*walkman*) and stroll out to “le parking” (the *parking lot*). Although all three of these truncated anglicisms are wholly English words, they do not carry the same meaning as their non-truncated English counterparts[12]. In sum, French inventions

[11] Rey-DeBove (1980) was the one of the only authors we encountered to provide other examples of this kind of anglicism. She points out that the French word “audience” in the sense of a *hearing*, has adopted the English meaning of *audience* like the crowd at a show. In addition, “alternative” in French means *alternate* as in “an alternate spelling”, whereas, when transformed into a semantic anglicism, it uses the English meaning of *alternative* (one of several possible solutions).

[12] Some other French modifications include the English word undergoing a grammatical class change in French (*fitness* (n.) in English becomes “faire fitness” (v.) in French), and singularization like with the English plural word *jeans* becoming the singular French word “un jeans”.

and modifications borrow an English word into French while leaving behind the English meaning.

3.1.6 MORPHOLOGICAL ANGLICISMS

Up to this point, the most prevalent categories of anglicisms in French relate to meaning and their transfer into French, yet another small category deserves brief mention. Although limited in scope, Grigg (1997), Spence (1989), and Trescases (1982) all attest to the existence of a morphological anglicism category in French. The most commonly cited morphological anglicism involves the suffixation of the English "-ing" morpheme (e.g., "brushing"). Another example of a morphological anglicism is the verb "lifter" where the French infinitive inflection "-er" is suffixed onto the English word *lift*. In this way we can regard the relatively rare morphological anglicisms as English words that undergo or cause some type of change word-internally when borrowed into French.

The six categories of anglicisms that have been explained in this section are summarized in Table 2 along with some examples.

Type	**Definition**	**Example**
Wholesale anglicism	A word or expression in English that is (usually) identical in both form and meaning to its usage in French.	- *weekend, muffin, bowling, fair play, sweat shirt, look* (n.), *Hi-Fi, steack, rosbif*
Direct translation	Borrowing an English object and its meaning and directly translating it into French.	-*gratte-ciel, haut parleur, chasseur de têtes, hors de la loi, effet de serre, nettoyage éthnique, facteur-risque*
Semantic anglicism	Borrowing an English word that is similar to an already existent French word (with a different meaning), and applying the English meaning to the French word.	- *réaliser* (in the English sense of *to realize*, not *to fulfill* or *to achieve* as in French) - *audience, alternative*
Hybrid	An English base word with added French elements.	- *surbooker, en live, top modèle*
French invention & modification	Borrowing an English word form into French without borrowing the English meaning.	- *tennisman, footing, un lifté/ un transplanté, slip* (n.) - *walk* (walkman), *straight pipe, un jeans*
Morphological anglicism	The addition of a morpheme that changes the meaning of the word through inflection and/or derivation.	- *brushing, forcing, lifter* (v.)

Table 2: Anglicism by category

4 RESEARCH QUESTIONS

As exemplified in the above review of previous research, frequency studies have touched on important factors such as anglicism use (tokens and types) in different varieties of French, and anglicisms in written language and in spoken language. However no single study to date has investigated a combination of the factors together in one body of research. The current study seeks to fill a gap in anglicism frequency research by addressing these shortcomings through the consolidation of numerous factors in the production of anglicisms in French. This research will explore three questions specifically:

15. Which language variety of French (France vs. Quebec) uses a higher total percentage of anglicisms?

16. Which language mode (written vs. spoken) is characterized by a higher frequency of anglicisms?

17. What is the effect of a type/token frequency distinction in analyzing anglicisms in French?

5 VARIABLE SELECTION AND CORPUS DESIGN

The objective of this study is to investigate the percentage of anglicisms in a representative corpus of the French language. Its design is based on a corpus of written and transcribed spoken data collected and compiled specifically for this study so as to take into consideration three major factor categories: *language variety* (French from Quebec, and French from France), *language mode* (written French, and spoken French), and the *token/type* distinction. In order to procure comparable oral and writing samples from French speakers from both France and Quebec, data from two television programs (Star Academy/Star Académie) and text from web logs (blogs) were collected, compiled and analyzed[13].

5.1 Language Variety

5.1.1 BACKGROUND

As mentioned above, this study will compare French from two distinct language varieties: French from France and from Quebec. The importance of studying more than one variety of French was highlighted in Mareschal (1992), and thus chosen as a factor for the present study. The reasons for which anglicisms are employed in each of the two language varieties is beyond the scope of this study; however, the authors concur that France and Quebec both borrow anglicisms for different reasons (Bouchard, 1999; Martel, 1991; Nadeau & Barlow, 2003). This topic will be revisited in this paper's "Results" section in light of the current findings and their relevance to the research questions posed.

5.1.2 CURRENT DESIGN

Because there are different reasons for borrowing and there exist different categories borrowed between users of France French and Quebec French, any study of anglicism frequency requires the separate consideration of the two language varieties. In the case of the current study, language variety will be defined as prototypical French written or spoken in two distinct regions: France and Quebec. The two subcategories France French and Quebec French are defined as the French produced by natives from each region. Natives, in turn, are people not only born in the region, but also residing in (and presumably participating in) the language community at the time of data collection.

[13] One general concern pertains to the importance of random sampling. The scope of this study involves investigating the effects of language variety and language type on anglicisms. This means that certain variables, such as age and gender, will not be accounted for. Yet a certain level of randomness still remains in the sampled television program participants and blog authors. For example, the television program is designed to select participants from various parts of the region. Moreover, the blog data will be completely random other than controlling for the authors' place of birth and residence and topic of discussion. The careful design control allows for this study to be easily replicable in the future by other researchers wishing to investigate the same or even alternate variables.

5.2 Language Mode

5.2.1 BACKGROUND

The second main variable under investigation is *language mode*. Research in corpus linguistics and discourse analysis has also revealed that certain linguistic features may vary depending on whether the data are gathered from spoken or written language (Tannen, 1982; Chafe, 1985; Louwerse, McCarthy, McNamara, & Graesser, 2004). In most sociolinguistic literature, for example, it is assumed that spoken language is less monitored than written language. The assumption can be taken one step further to apply to anglicisms in that written language may contain fewer anglicisms if they are seen as "unwanted/undesirable". On the other hand, as discussed earlier, previous research has observed different reasons for the use of anglicisms across different language varieties. In regions where English borrowings are viewed as prestigious (i.e., France), the more deliberate monitoring of written language could produce a higher percentage of anglicisms than spoken language due to the author wanting to assert a certain social status.

In the end, the reasons for anglicism use cross-regionally require further investigation, and are beyond the scope of the present study. Chafe and Tannen (1987), in a comprehensive overview of research investigating the differences in written and spoken language, conclude that "different conditions of production as well as different intended uses foster the creation of different kinds of language" (p. 390). It is therefore safe to assume that linguistic borrowings, such as anglicisms in French, may not be exempt from the type of variation that affects spoken and written language modes.

5.2.2 CURRENT DESIGN: WRITTEN DATA

In order to procure writing samples from French speakers, text produced (and possibly edited) by authors of various publicly posted web logs (blogs) was collected from Internet websites. For both varieties of French, the content of the blogs remained within the realm of everyday living and family matters (e.g., advice on shopping, job search stories, trouble with spouses and children, etc.).

The question arises, of course, as to whether the authors of these blogs are speakers of the language variety of French from France or from Quebec, a factor which was controlled for in a number of ways. One way of ensuring authorship for the target language variety was by refining Internet searches for blogs to the region-specific domain extensions (i.e., blogger.fr, or blogspot.ca). Although many French and Quebec authors use blog websites with ".com" extensions, which obscure the authors' origins, this step was helpful in immediately eliminating many English language blogs. Next, the principle means of confirming the origin and region of residence of the blog authors was via direct personal communication. This communication, in the form of an email, indirectly inquired about the authors' birth location and region of current residence (either Quebec or France). Once authors had replied, blogs were then chosen based on whether or not replies were consistent with target data. Finally, once the above two methods had been employed, the actual content of the texts, read while reviewing the data, provided valid information about the authors. This last step was especially useful in identifying and eliminating authors who had spent an extended period of time living or traveling in regions where they may have been heavily influenced by the English language (thus potentially skewing the results to reflect a higher frequency of anglicisms in their blogs).

In addition to authorship, the date of the data collected from the blog websites was also taken into consideration, that is, blog postings prior to fall 2008 were not considered for either varieties of French. This ensured that the texts used for data collection were both methodologically equivalent across language varieties as well as quite recent. Finally, the

total number of blog sources and words subjected to analysis remained equal across both language varieties.

5.2.3 CURRENT DESIGN: SPOKEN DATA

The corpus design for the spoken data variable was configured somewhat differently. Data for spoken French for each variety were gathered from the daily lives of French speakers in two reality television programs: *Star Academy* (in France, *www.tf1.fr/star-academy*) and *Star Académie* (in Quebec, *www.staracademie.ca*), two versions of the exact same program (same concept, and format) with the same target audience (for the sake of comparability)[14].

Since the study of anglicisms necessarily involves English words in French, it was important for the sake of internal validity to choose programs that would not be inherently subject to either more or less anglicisms than in typical spontaneous speech. Scripted sitcoms and dramas, for instance, give producers and directors the opportunity to potentially add or delete anglicisms in the script based on the target audience, political standpoints, or other TV network agendas. On the other hand, unscripted reality-format programs analyzed in the present study provided the ideal medium for spontaneous, unmonitored speech and allowed for spontaneous language production in the widest possible range of topics. Moreover, the original series *Star Academy,* produced by European company Endemol, was first aired in France in 2001 ("Star Academy", 2009) and was adopted in Quebec with the same format and successfully broadcasted as *Star Académie* by 2003. Consequently, the choice of a television program that is not a spin-off of an English concept helped control once again for any unnecessary outside influence from Anglophone culture.

Another important factor in the selection of an appropriate television program is the recency of the data. Previous studies of anglicisms in French are criticized here as using outdated data (e.g., Mareschal, 1992, who uses data from the late 1970s). Current data samples are a unique feature of this study and are deemed important as anglicisms are a particularly dynamic part of language (Nadeau & Barlow, 2005). Collecting and using recent data assured that the anglicisms found in the data are representative of current day spoken and written anglicisms. Finally, again for comparability sake, language recorded from each of the two television programs represented a similar word count for the two varieties under investigation.

5.3 Tokens and Types

The final factor group of the current study is the distinction between tokens and types of anglicisms. An anglicism token is defined by counting each and every anglicism as a separate occurrence, whereas an anglicism type only counts anglicisms from a new word family (e.g., *blog*, *blogeur*, *blogosphère* constitute three tokens but only one type; the same holds for *weekend*, *weekend*, and *weekend* – three tokens, one type). This token-type aspect of quantification is a particularly important consideration in anglicism frequency counts. Token counts give a sense of density, of whether one particular anglicism is highly used and thus artificially inflating the number of anglicisms, while type counts show the degree of variation/diversity of the different anglicisms in the corpus.

[14] The difference in orthography between the two television show titles is curious and deserves mention at this juncture. *Star Academy* in France is completely English, a wholesale anglicism, while the Quebec version of the show *Star Académie* appears to attempt a frenchification of the title. Ironically, however, in doing so the title still remains an anglicism but of the direct translation category. Indeed, "Star" aside, in order for the title to be truly French, the title would need to read "Académie des Stars". Thus, both titles from both varieties of French can be considered anglicisms, but of different categories.

Recall the numbers and percentages discussed in previous studies (see Table 1). The percentage of anglicism occurrences in Forgue (1986) and Mareschal (1992) were all significantly lower for types than for tokens, while the opposite was true for Théoret (1991). It is clear here that if research fails to consider both tokens and types side by side, the reader may only see one aspect of the data. It is therefore imperative for any investigation of anglicism percentages to look not only at the token frequency, but also the frequency of anglicism types in order to get a full understanding of the linguistic feature's true frequency and distribution, and not be mislead to draw under informed or misguided conclusions from the data.

6 DATA MANIPULATION AND ANALYSIS

In order to accurately measure anglicisms, a triangulation of methods and instruments was devised. Firstly, the Microsoft Word spell-check feature in French helped visually locate unrecognized or "misspelled" (i.e., borrowed) words. Once identified as an anglicism, the "find" search feature in Word was employed. This feature proved useful, once a particular anglicism had been identified, in locating any and all other recurring instances of the same anglicism in the corpus.

The next step in identifying anglicisms in the corpus was by referencing the table of definitions in Table 2. A final method for locating anglicisms in the corpus was to consult dictionaries/lists of anglicisms in French (i.e., Forest, 2006; Hofler, 1982; Laurin, 2004; and Rey-DeBove & Gagnon, 1980) and cross-reference the information with a French dictionary (Larousse, 2004).

The procedure for this study was relatively straightforward. The first step involved the data collection itself. For the *Star Academy*/*Star Académie* television programs, this involved careful watching and listening to each recorded video file in conjunction with the task of transcribing the files into text. For the blog entries, a certain amount of "clean-up" was necessary in order to render the text into an analyzable corpus. This involved taking out pictures and photos; deleting "non-text" symbols such as smiles and extra punctuation; deleting sounds and verbalized emotions (haha, hihi, lol, pfffft!); and deleting outside references or citations from any language (i.e., poems, quotes, movie dialogues, song lyrics, "top 10" lists, recipes, etc.). These measures were taken in order to avoid distorting the data's representative portrayal of anglicisms. Ultimately, for both language types it was necessary to rectify the word count from the data sources to ensure equivalence across language variety.

The second step involved analyzing the data and identifying all instances of anglicisms in the corpus. As described above, anglicisms were identified through a triangulation of methods and instruments: Microsoft Word French spell-check and find feature, a summary chart of definitions from the literature, and dictionaries of anglicisms. This analysis was conducted keeping in mind the three independent variables, language type, language mode (oral/written), and the token/type distinction.

6.1 Results and Implications

The analysis of the data on both written and spoken anglicism tokens and types in France and Quebec yielded intriguing results. The raw numbers and percentages from the frequency counts, organized in vertical columns by language variety and language mode, and in horizontal columns by anglicism tokens and types, are presented in Table 3.

	FF WT	FF SP	FF total	QF WT	QF SP	QF total	Total
Token n=	85	69	**154**	65	182	**247**	**401**
Token %	0.84	0.67	**0.75**	0.64	1.82	**1.23**	**0.99**
Type n=	52	30	**82**	43	61	**104**	**186**
Type %	0.51	0.29	**0.40**	0.43	0.61	**0.52**	**0.46**

Table 3: Anglicism frequency totals

Firstly, in looking at the rightmost column "Total" and reading across the first row, one can observe the total number of anglicism tokens in a corpus (written and spoken) of over 40,000 words to be just over four hundred. This translates into just under one percent (second row, 0.99%). As for total number of anglicism types, the number decreases by more than half. That is to say, only 0.46% of the word types in the corpus are anglicism word types (n=186). The interpretation of these totals suggests that the overall percentage of anglicisms in French, when compared with the findings from the previous studies discussed at the start of this paper (see Table 1), has not dramatically increased over the last two decades, as it remains around 1% of the total words analyzed.

Next, the analysis yielded data relevant to anglicism differences across language varieties. The relationship between the total anglicisms in these two language varieties ("FF total" and "QF total" columns) becomes more apparent in Table 4.

	FF	QF
Token	154 (0.75%)	**247** (1.20%)
Type	82 (0.40%)	**104** (0.50%)

Table 4: Anglicisms by language variety

Recall the first research question posed at the outset of this paper: *Which language variety of French uses a higher total percentage of anglicisms?* In Table 4 it becomes immediately clear that the variety of French spoken and written in Quebec contains more anglicisms than French from France. The number spread, however is not equal between tokens and types. When looking at the difference in anglicism percentages across the "Token" row, for example, QF contains more anglicisms than FF by almost half a percent (0.45%). On the other hand, however, the difference in the amount of anglicism types used in QF versus the amount in FF is reduced to only a tenth of a percent (0.10%).

These results demonstrate how the Quebec language variety of French uses overall more anglicisms than French in France. However, several caveats are in order regarding additional factors that may have shaped these results. Firstly, due to the difficulty in identification and general rarity of the categories, both morphological anglicisms and semantic anglicisms were omitted from the current data. This resulted in an analysis of only four of the six anglicism categories described earlier in this paper. In order to get the most precise view of anglicism frequency in a language, future research would have to include all six categories.

In addition, the size of the corpus (over forty thousand words total) could have possibly weighted the anglicism occurrences more heavily than in a larger corpus. Although in theory, if a corpus is representative of the population it portrays (which this study has endeavored to be, see previous section on the study's design), increasing the data set should not alter results. And yet, the possibility still remains that a higher or lower percentage of anglicisms in one language variety or another could emerge as a result of a larger corpus.

Finally, there remains the fact that certain individual speakers (in the TV data) or authors (in the blog data) may have possessed a certain affinity for or aversion to anglicization. For

example, it was observed by one of the current authors (though not formally documented in the data) that one particular speaker from Québec's *Star Académie* produced a substantially higher number of anglicisms than the other speakers on the show. It is even possible, though harder to observe, that a language user may consciously avoid the use of anglicisms. Regardless of reasons, these individual behaviors would undoubtedly have a smaller effect on the data in a larger corpus.

The answer to the second research question, *Will one language mode yield a higher frequency of anglicisms than the other?*, proves less straight forward than the first (see Table 5). When considering the overall picture of language mode (written versus spoken) it looks as if anglicisms are considerably more prevalent in spoken language (n=251) than in written French text (n=150) when considering word tokens. On the other hand, when looking at anglicism word types, the opposite is true. Ninety-five written anglicism types prove slightly higher than the 91 counted spoken types.

	Written	**Spoken**
Token	150	251
Type	95	91

Table 5: Anglicism totals by language mode

If stopping here, the implications of these results prove mystifying. And yet, when expanding the view of language mode by adding the variable of language variety, a visible pattern emerges (Table 6).

		Written	**Spoken**
FF	(*token*)	**85**	69
	(*type*)	**52**	30
QF	(*token*)	65	**182**
	(*type*)	43	**61**

Table 6: Anglicisms by language mode and variety

In Table 6, it becomes evident (for both tokens and types) that anglicisms are more common in written French in the language variety from France, and that conversely, anglicism use in the Quebec variety of French appears higher in spoken language. The fact that France employs more anglicisms in writing while Quebec uses more in speech could very possibly be attributed to the different reasons each of these language varieties borrows from English. Research on French specifically has found that language users from France use anglicisms in French for different reasons than French language users in Quebec.

Indeed, a vast body of literature has been dedicated to distinguishing the reasons for anglicism use in France and in Quebec (see Timmins, 1995; Théoret, 1991; Nadeau & Barlow, 2003). For example, it has been argued that French speakers from France tend to employ anglicisms due to their "fondness" of American culture (Timmins, 1995) and due to the historical prestige status these English borrowings hold as they have always been associated with higher bourgeoisie social groups and good taste (Bouchard, 1999).

Conversely, in Quebec, working-class ex-peasants during the Industrial Revolution began to resent the dominant upper class English society and their novel words/concepts (i.e., le *boss*, le *shop*, le *foreman*, le *drill*, etc.) causing these and other anglicisms to become stigmatized to the point where French speakers looked to replace them with a French form, correct or not (Bouchard, 1999; Forest, 2006; Timmins, 1995). This history, combined with a

direct contact with a predominantly English-speaking continent, and continual exposure to English culture through sports, work, and brands, has shaped the nature of and reasons for anglicisms in present-day Quebec French (Nadeau & Barlow, 2003; Timmins, 1995).

These reasons, of course, are not the only explanations for anglicisms in French, nor have the results of the present study served to definitively propose these explanations as facts. What is clear from the data, however, is that in order to explain why FF contains more anglicisms in written language and why QF has more anglicisms in spoken French, additional research on the relationship between language mode and reasons for borrowing is required.

A final word regarding the third research question posed at the outset of this paper: *Is the distinction between anglicism type frequency and token frequency relevant*? The response to this question manifests itself in the context of the results just discussed. In the first research question pertaining to language variety, for instance, the comparison of token word frequency with type word frequency revealed differences in the distances between the two language types (see Table 4, where we show that the gap in numbers between QF and FF was much smaller for types than for tokens). Moreover, while looking at token frequency gives a general sense of a linguistic feature's (i.e., anglicisms) pervasiveness and density in the data, type frequency reveals information regarding the diversity (new unique words as opposed to the same word repeating over and over) of that feature in the corpus. That is to say that in the study of anglicism frequency, and in any linguistics frequency study for that matter, it is imperative to take both token and type frequency into consideration. As exemplified in the current data, different results emerge based on whether tokens or types are analyzed. Without simultaneously considering both of these two measures, any results collected from the data will be portraying only part of the picture.

7 CONCLUDING REMARKS

This study sought to gather and interpret empirical data regarding the number of anglicisms in a French corpus, in which language mode, language variety, and the distinction between token frequency and type frequency were taken into consideration. Three research questions guided the study, and careful design and procedures were adopted in order to sufficiently answer these questions.

One major contribution of this experiment to cross-linguistic studies is the quantification of a familiar phenomenon (borrowings from English to French) in a methodologically sound way. Théoret (1991), for example, laments the difficulties of speaking about anglicisms in Quebec. Since there are not a large number of objective studies on the subject, people appoint themselves the right to speak about and pass definitive judgments on the topic of anglicisms as if the fact of knowing *how* to speak a language allowed them to analyze it in a scientific and impartial manner (p. 79).

The fact that the data collected was recent also brings a substantial amount of validity to the study. This point is especially important to control for when dealing with a dynamic language element such as anglicisms (Clyne, 2003). Furthermore, by using reality television shows (spontaneous and unmonitored speech) and blog texts (freely composed by amateur authors), as opposed to edited television programs or newspaper/magazine articles, the corpus compiled and analyzed derives from "authentic" (less monitored) spoken/written contexts.

An additional contribution of this study pertains to the use of both written and spoken data. Many studies to date (see Cerquiglini, 1991; Mareschal, 1992; Théoret, 1991) have investigated anglicisms in either written or spoken corpora but not both. The significance of incorporating both language modes was discussed earlier in the "Variable Selection and Corpus Design" section of this paper, and the results from the current study have indeed

confirmed anglicisms as a linguistic feature susceptible to written language/spoken language differences.

Similarly, another unique aspect of this study is the use of Internet blog data for the written language variety. By using written data from blogs, it was assumed that the level of formality would fall somewhere between formal writing and informal speech. One of the most widely studied variables in sociolinguistics over the past several decades remains the formality of speech. Since Labov's (1966) famous *fourth floor* study on formality in various situations and contexts, many subsequent studies by researchers in linguistics have sought to uncover the relationship between particular linguistic features and language formality (e.g., Cardoso, 1999, 2003, 2007; Lin, 2003; Major, 2001, 2004). This study has observed a difference between written and spoken language, and yet it remains to be determined whether this difference can be attributed to a potential more careful monitoring of language in either of the two modes investigated.

The goal of this study was to use authentic up-to-date informal writing and speech as tools to investigate the percentages of anglicisms in two varieties of French. Furthermore, compiling a new and unique corpus of current written and spoken data creates an invaluable opportunity for future research on different variables pertaining to anglicisms or any other range of linguistic features in spoken and/or written language. In the end, the question still remains as to at what point the presence of English in everyday French would be considered "contamination" or "infiltration" of the language. Yet, in undertaking such a study, it is hoped that the results will help fill the gaps in the overall body of knowledge in the field of borrowings especially concerning the frequency of anglicism use so that this empirical data can now help substantiate those good-humored finger-pointing bar discussions that claim "your French variety uses more anglicisms than mine".

8 REFERENCES

Bouchard C. (1999). *On n'emprunte qu'aux riches: La valeur sociolinguistique et symbolique des emprunts*. Montreal: Fides.

Cardoso W. (1999). A quantitative analysis of word-final /r/-deletion in Brazilian Portuguese. *Linguistica Atlantica* 21, p. 13-52.

Cardoso W. (2003). *Topics in the phonology of Picard.* PhD Thesis, McGill University. Published by the McGill Working Papers in Linguistics.

Cardoso W. (2007). «The variable development of English word-final stops by Brazilian Portuguese speakers: A stochastic optimality theoretic account.» *Language Variation and Change* 19, p. 1-30.

Cerquiglini B. (1991). «Le point de vue du français européen.» In P. Martel and H. Cajolet- Laganière (eds.), *Actes du colloque sur les anglicismes et leur traitement lexicographique: Communications, discussions et syntheses: Magog du 24 Au 27 Septembre 1991*. Quebec: Gouvernement du Quebec, Office de la langue francaise. p. 297-301.

Chafe W. (1985). «Linguistic differences produced by differences between speaking and writing.» In D. R. Olson, N. Torrance and A. Hildyard (eds.), *Literacy, language, and learning: The nature and consequences of reading and writing*. New York: Cambridge University Press. p. 105-123.

Chafe W. and Tannen D. (1987). «The relation between written and spoken language.» *Annual Review of Anthropology, 16*, p. 383-407.

Clyne M. (2003). *Dynamics of language contact: English and immigrant languages*. Cambridge: Cambridge University Press.

Dictionnaire du français plus: A l'usage des francophones d'amerique (1988). Montreal: Centre educatif et culturel.

Etiemble R. (1964). *Parlez-vous franglais?* Paris: Gallimard.

Forest J. (2006). *Les anglicismes de la vie quotidienne des Quebecois: Essai*. Montreal: Triptyque.

Forgue G. J. (1986). «English loan words in French today.» *Journal of English Linguistics*, *19*(2), p. 285-294.
Grigg P. (1997). «Toubon or not Toubon: The influence of the English language in contemporary France.» *English Studies, 78*(4), p. 68-384.
Guiraud P. (1965). *Les mots étrangers: Que Sais-Je?* Paris: Presses Universitaires de France.
Hofler M. (1982). *Dictionnaire des anglicismes*. Paris: Larousse.
Labov W. (1966). *The social stratification of English in New York City.* Washington: Center for Applied Linguistics.
Laurin J. (2004). *Les américanismes: 1200 mots ou expressions made in USA*. Montreal: Editions de l'Homme.
Le petit Larousse illustré: en couleurs (100th ed.). (2004). Paris: Larousse.
Le Prat G. (ed.) (1980). *Dictionnaire de franglais: Plus de 850 mots et locutions de langue anglais couramment utilizes dans les medias, la conversation ou la correspondance française d'aujourd'hui et leur traduction en français.* Paris: Guy Le Prat.
Lin Y.-H. (2003). «Interphonology variability: sociolinguistic factors affecting L2 simplification strategies.» *Applied Linguistics* 24 (4), p. 439-464.
Louwerse M. M., McCarthy P. M., McNamara D. S. & Graesser A. C. (2004). «Variation in language and cohesion across written and spoken registers.» In K. Forbus, D. Gentner and T. Regier (eds.), *Proceedings of the twenty-sixth annual conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum. p. 843-848.
Major R. (2001). «Linguistic explanations for second language phonological systems. [Chapter 2].» *Foreign Accent. The Ontogeny and Phylogeny of Second Language Phonology*. New Jersey: Lawrence Erlbaum Associates.
Major R. (2004). «Gender and stylistic variation in second language phonology.» *Language Variation and change* 16 (3), p. 169-188.
Mareschal G. (1992). «L'influence comparée de l'anglais sur le français dans différentes aires géographiques francophones.» *Journal of the Canadian Association of Applied Linguistics,* 14(2), p. 107-120.
Martel P. (1991). «Conference opening.» In P. Martel & H. Cajolet- Laganière (eds.), *Actes du colloque sur les anglicismes et leur traitement lexicographique: Communications, discussions et syntheses: Magog du 24 Au 27 Septembre 1991*. Quebec: Gouvernement du Quebec, Office de la langue francaise. p. 297-301.
Mougeon R. and Beniak E. (1991). *Linguistic consequences of language contact and restriction: The case of French in Ontario, Canada*. Oxford: Clarendon Press.
Nadeau J. and Barlow J. (2003). *Sixty million Frenchmen can't be wrong: Why we love France but not the French*. Naperville, IL: Sourcebooks, Inc.
Nadeau J. and Barlow J. (2006). *The Story of French*. New York: St. Martin's Press.
Picone M. D. (1996). *Anglicisms, neologisms and dynamic French*. Amsterdam: John Benjamins Publishing.
Rey-DeBove J. and Gagnon G. (eds.). (1980). *Dictionnaire des anglicismes: Les mots anglais et americains en francais* (pp. *v - xvi*). Paris: Robert.
Robert P. (2005). *Le nouveau petit robert: Dictionnaire alphabetique et analogique de la langue francaise*. Paris: Dictionnaires Le Robert.
Rifelj C. (1996). «False friends or true?: Semantic anglicisms in France today.» *The French Review, 69*(3), p. 409-416.
Spence N. (1989). «Qu'est-ce qu'un anglicisme?» *Revue de Linguistique Romane, 53*, p. 323- 334.
Tannen D. (1982). «Oral and literate strategies in spoken and written narratives.» *Language, 58*(1), p. 1-21.
Théoret M. (1991). «La situation des anglicismes au Québec.» In P. Martel & H. Cajolet- Laganière (eds.), *Actes du colloque sur les anglicismes et leur traitement lexicographique: Communications, discussions et syntheses: Magog du 24 Au 27 Septembre 1991*. Quebec: Gouvernement du Quebec, Office de la langue francaise. p. 79-92.

Thogmartin C. (1984). «Some 'English' words in French.» *French Review, 57*(4), p. 447-455.
Timmins S. (1995). *French fun: The real spoken language of Québec*. Toronto: Wiley.
Trescases P. (1982). *Le Franglais vingt ans après: Langue et société*. Montréal/Toronto: Guérin.
Ullmann S. D. (1947). «Anglicisms in French: Notes on their chronology, range, and reception.» *Publications of the Modern Language Association of America, 62*(4), p. 1153-1177.
Villers M. D. (1988). *Multidictionnaire des difficultes de la langue francaise*. Montreal: Editions Quebec/Amerique.
Weinreich U. (1974). *Languages in contact: Findings and problems* (8th ed.). The Hague: Mouton.
Wise H. (1997). *The vocabulary of modern French: Origins, structure and function*. London: Routledge.

UN CORPUS ANTILLAIS D'APPRENANTS DE FRANÇAIS

Régis Kawecki
Laboratoire HCTI – Université de Bretagne-Sud
Centre d'apprentissage des langues – Université des Indes Occidentales – (Trinité et Tobago)

La linguistique de corpus est un acteur clé du domaine de l'apprentissage et de l'enseignement des langues. Les premières études ont été effectuées sur des corpus de locuteurs natifs. Ce n'est que plus récemment que des corpus d'apprenants se sont constitués à partir de productions d'étudiants en langues étrangères. D'abord presque exclusivement tournée vers l'enseignement/apprentissage de l'anglais langue étrangère, la linguistique de corpus s'intéresse désormais aux autres langues mais sans encore atteindre l'importance des corpus réalisés pour l'anglais. Les corpus d'apprenants peuvent apporter une aide précieuse à l'enseignement des langues étrangères car ils sont à la base du repérage des problèmes linguistiques les plus récurrents que rencontre une population d'apprenants particulière dans son apprentissage d'une langue donnée. Ce repérage permet ensuite d'adapter l'approche pédagogique aux difficultés spécifiques des apprenants en développant du matériel de classe supplémentaire distribués en complément de la méthode utilisée, celle-ci étant bien souvent par trop générale dans son approche. De tels corpus peuvent également être mis à la disposition des étudiants pour un apprentissage en autonomie. Cette présentation décrit le corpus d'apprenants de français qui se constitue au Centre d'apprentissage des langues de l'université des Indes occidentales de Trinité et Tobago dans les Antilles anglophones. Il s'agit d'une collection de courts écrits réalisés durant plusieurs semestres de classe par les étudiants du centre apprenant le français, des niveaux débutants jusqu'à intermédiaires avancés. Les premières analyses effectuées montrent que les apprenants trinidadiens de français diffèrent de façon significative d'autres populations d'étudiants de langue anglaise.

La linguistique de corpus a d'abord investi les domaines de la lexicographie et de la linguistique descriptive. Les recherches ont été réalisées à partir de corpus de locuteurs natifs, écrits ou oraux (Biber, 1998). Le *International Corpus of English* (ICE), hébergé par l'*University College* de Londres et utilisé pour des études comparatives sur différentes variétés d'anglais parlées dans le monde, est un cas représentatif de ce type de corpus.
La linguistique de corpus s'est ensuite fait une place dans l'enseignement/apprentissage des langues. De l'apprentissage de l'anglais sur objectifs spécifiques à partir de corpus de locuteurs natifs jusqu'à l'apparition de méthodes en langue étrangère validée par le recours aux corpus d'apprenants, la linguistique de corpus a investi l'enseignement/apprentissage des langues, à tel point que « les corpus font maintenant partie des outils que de plus en plus de professeurs s'attendent à utiliser en classe (Sinclair, 2004) ».
Cette présentation rend compte des premiers résultats d'une thèse de doctorat réalisée par l'auteur de ces lignes et dirigée par le professeur Geoffrey Williams de l'université de Bretagne-Sud. La recherche menée cherche à articuler certains acquis en linguistique de corpus à l'enseignement/apprentissage du français langue étrangère. Elle a pour finalité

essentielle de mettre à la disposition des apprenants des outils d'analyse de leur production langagière qui permettent de les sensibiliser davantage aux problèmes que pose l'apprentissage d'une langue étrangère. Cette recherche a eu lieu dans le cadre du Centre d'apprentissage des langues de l'université des Indes occidentales de Trinité et Tobago dans les Antilles anglophones.

1 LE CENTRE D'APPRENTISSAGE DES LANGUES

Le Centre d'apprentissage des langues (CAL) de l'université des Indes occidentales de Trinité et Tobago se consacre à l'enseignement des langues étrangères pour les étudiants de l'université ne se spécialisant pas en langue(s) mais désireux de quitter l'université avec une compétence langagière supplémentaire. Les cours sont également ouverts aux personnes externes à l'institution. Les deux langues les plus demandées sont l'espagnol et le français mais sont également proposés l'arabe, le mandarin, l'allemand, le hindi, l'italien, le japonais, le portugais, le yoruba et l'anglais en tant que langue étrangère. Le centre offre trois niveaux d'enseignement ce qui correspond à six semestres d'étude. Un étudiant de première année peut donc ajouter une langue étrangère à son bagage pendant tout le temps de sa licence. Dans la pratique, seuls l'espagnol et le français offrent un cursus complet, les autres langues n'allant guère au-delà de trois semestres. L'approche pédagogique est communicative pour l'ensemble des quatre compétences. Il s'agit avant tout de communiquer dans une langue standard à propos de sujets personnels ou d'intérêt général dans des situations de vie quotidienne. Les apprenants du centre atteignent en fin de parcours un niveau équivalent au B1/B2 du Cadre européen commun de référence pour les langues (CECR), cette fourchette s'expliquant par leur plus ou moins grand investissement dans cet apprentissage. Dans la nomenclature de l'ACTFL, *the American Council on the Teaching of Foreign Languages*, ces niveaux correspondent à ceux de *Intermediate Mid/High*.

2 CORPUS ET ENSEIGNEMENT DES LANGUES

Les corpus de locuteurs natifs ont été les premiers à faire l'objet d'une utilisation en cours, notamment dans l'enseignement des langues sur objectifs spécifiques ou langues de spécialité telles que l'anglais scientifique ou le français des affaires. Des étudiants en biologie, qu'ils soient natifs ou pas, peuvent ainsi avoir accès à un corpus d'articles publiés dans leur domaine et s'initier de cette façon, à l'aide de cet outil, à la présentation de leurs travaux de recherche (Partington, 1998) dans l'optique d'une publication dans les revues spécialisées.

Les études en traduction font également largement appel aux corpus de locuteurs natifs souvent dans ces mêmes domaines de spécialité. Les corpus peuvent être parallèles, ce qui permet de visualiser sur une même page d'écran l'original et sa traduction. La fréquentation répétée de tels corpus permet aux futurs traducteurs de développer des savoir-faire qui seraient plus laborieux à acquérir sans l'aide de cet outil.

Dans les années 50, il faut signaler une première significative : *Le français fondamental* (Gougenheim, 1956). En se basant sur un corpus oral enregistré ainsi que sur la langue des journaux, ce travail pionnier, sous la direction de G. Gougenheim et de P. Rivenc, a permis l'élaboration d'une liste des mots et indications grammaticales les plus usités en français et a longtemps servi à la création de méthodes d'enseignement du français langue étrangère plus adaptées au discours oral tel qu'il se pratiquait dans la vie de tous les jours.

3 CORPUS D'APPRENANTS

Ce n'est que plus récemment que certains éditeurs ont commencé à rassembler des productions d'étudiants en langue seconde ou étrangère afin de bâtir des corpus d'apprenants. Ceux-ci ont permis l'élaboration de dictionnaires et de méthodes mieux adaptées aux populations concernées et aux contenus d'apprentissage. Ces corpus ont d'abord été développés pour l'enseignement de l'anglais langue seconde ou étrangère. Deux exemples bien connus de corpus d'apprenants en langue anglaise sont le *Cambridge Learner Corpus* ainsi que le *Longman Learners' Corpus*.

Dans une optique non commerciale, certains universitaires se sont également lancés dans la compilation de corpus d'apprenants. C'est par exemple le cas de Sylviane Granger, de l'université de Louvain la Neuve en Belgique, qui est à l'origine du *International Corpus of Learner English* (ICLE), une collection d'écrits d'étudiants en anglais, de niveau avancé, originaires de différents pays (Granger, 1998).

De nombreux corpus d'apprenants d'anglais ont depuis vu le jour. Ces collections cherchent à rendre compte des difficultés spécifiques que rencontrent différentes populations d'étudiants dans leur apprentissage de la langue anglaise. Il existe, pour n'en citer que quelques-uns, des corpus d'apprenants chinois d'anglais à Taiwan comme à Hong Kong. L'université des sciences et technologies de Hong Kong est par exemple à l'origine d'un corpus conséquent de ce type établi par John Milton (1994). Comme les différentes variétés d'anglais langue maternelle peuvent être comparées entre elles à l'aide de corpus de locuteurs natifs (comme pour le ICE), il est désormais également possible de mener des études contrastives sur les différents corpus d'apprenants d'une même langue cible tel que l'anglais.

Pour les autres langues, des corpus d'apprenants voient désormais le jour mais restent de taille relativement modeste. Parmi les plus récents, il y a par exemple le *International Corpus of Learner Finnish* constitué à l'université de Oulo en Finlande sous la direction de Jantunen.

Les corpus d'apprenants de français langue étrangère restent quant à eux peu nombreux et de taille relativement réduite.

Force est de constater que ces corpus d'apprenants sont le plus souvent des collections de productions d'allophones qui ont un bon niveau dans la langue cible. Les corpus d'apprenants débutants et intermédiaires, par contre, correspondants aux niveaux A1 à B2 (CECR) ou bien *Novice Low* à *Intermediate High* (ACTFL), sont beaucoup plus rares.

4 CORPUS D'APPRENANTS ET ENSEIGNEMENT DES LANGUES

Les corpus d'apprenants, « ces collections digitalisées de textes produits par des étudiants de langue seconde ou étrangère (Granger, 1998) », permettent deux types d'avancées en enseignement des langues.

Au niveau théorique, ils s'avèrent particulièrement utiles pour comprendre les processus à l'œuvre dans l'acquisition d'une langue seconde ou étrangère. Ils contribuent notamment à l'étude du concept d'interlangue, appelé encore grammaire de l'apprenant ou bien encore grammaire provisoire. Le concept d'interlangue dans le domaine de l'acquisition des langues s'applique au système linguistique, par essence incomplet et imparfait, mis en place par les locuteurs non-natifs dans leur tentative de s'expliquer le fonctionnement de la langue qu'ils sont en train de découvrir. Hanzeli (1975) parle de « The provisional grammar that each learner develops and refines in the very process of learning[15] ». Les apprenants, en partie de façon automatique, parfois par généralisations excessives, se créent un jeu de règles

[15] La grammaire provisoire que chaque apprenant développe et qu'il perfectionne par le fait même d'apprendre.

personnelles qui leur permet, au niveau d'apprentissage qui est le leur, d'organiser la matière linguistique qu'ils découvrent et la manière originale qu'elle a de rendre compte du monde réel. Cette interlangue est par définition changeante. Elle évolue parallèlement à l'enseignement dispensé ou à l'auto-apprentissage réalisé. Une analyse longitudinale de corpus d'apprenants permet d'en comprendre les différentes étapes (Barlow, 2005).

Au niveau des pratiques de classe, le recours aux corpus d'apprenants permet une plus grande efficacité de la pédagogie qui sous-tend l'enseignement des langues étrangères. L'analyse de corpus d'apprenants doit pouvoir déboucher au minimum sur la création de matériel pédagogique supplémentaire centré sur les problèmes particuliers rencontrés par une population spécifique d'apprenants. Ce complément viendra corriger la méthode utilisé, celle-ci ayant le plus souvent une approche par trop générale de la langue. Pour des raisons de rentabilité commerciale essentiellement, les livres vendus aux étudiants s'adressent à un apprenant « mondialisé » que l'on ne rencontre nulle part. Les apprenants chinois d'anglais langue étrangère par exemple font face à des difficultés d'apprentissage qui n'ont rien à voir avec celles auxquelles se heurtent des apprenants français d'anglais. Leurs interlangues ne sont certainement pas identiques ne serait-ce que parce que des langues maternelles aussi éloignées que le mandarin et le français interfèrent de façon très différente dans le processus d'acquisition de l'anglais. Au mieux, les corpus d'apprenants peuvent être à la base de méthodes d'enseignement ciblées à destination de publics spécifiques. Cette possibilité n'est pas la plus répandue, loin s'en faut, et les quelques adaptations de méthodes génériques, réalisées en Afrique par exemple, s'attachent davantage aux aspects culturels que linguistiques.

5 CORPUS D'APPRENANTS ET INTERLANGUE

Les recherches en interlangue peuvent s'articuler selon la dichotomie saussurienne (Saussure, 1964) de synchronie/diachronie. Les corpus d'apprenants permettent l'étude synchronique ou diachronique des productions d'une population particulière d'apprenants. Une analyse synchronique rend compte des spécificités linguistiques caractéristiques de la façon d'écrire ou de s'exprimer d'une communauté d'étudiants à un moment donné de leur apprentissage. Les corpus d'apprenants permettent également l'analyse étalée dans le temps de la production langagière des apprenants. Les chercheurs peuvent donc rendre compte des caractéristiques lexicales ou morpho-syntaxiques typiques d'un groupe d'étudiants aux différents paliers atteints par ceux-ci, tels qu'ils sont décrits par le Cadre européen commun de référence pour les langues (CECR) par exemple, ou bien encore pointer du doigt les évolutions constatées tout au long de leur étude de la langue étrangère ou seconde, ce qui relève notamment du domaine de l'acquisition des langues.

Les spécificités de l'interlangue d'apprenants dépendent généralement des processus généraux à l'œuvre dans l'acquisition des langues comme (Ellis R. et Barkhuizen G., 2005) :

18. l'intrusion de la langue maternelle (L1),
19. l'influence de la méthode utilisée : le « formulaic parroting[16] » de l'ACTFL,
20. l'influence d'un registre de langue dominant qui peut avoir pour conséquence une langue écrite proche de l'oral : utilisation répétée de « ça » à l'écrit par exemple,
21. la généralisation intralangue: des généralisations abusives d'aspects de la langue étudiée,
22. les étapes du développement de l'interlangue : une nouvelle donnée linguistique est d'abord comprise et acceptée de façon théorique par l'apprenant et ce n'est qu'avec du temps et de la pratique qu'elle finit par passer du stade de la

[16] Répétitions telles quelles.

compréhension passive à celui de l'expression active pour apparaître dans le discours et les écrits de l'étudiant,

23. les stratégies d'adaptation aux difficultés d'apprentissage d'une langue étrangère comme l'évitement (Odlin, 1989).

S'ajoute à cette liste, dans certaines régions du monde :

24. l'intrusion d'une langue seconde/étrangère (L2) particulièrement présente telle que l'espagnol dans le cas de Trinité et Tobago.

L'interprétation et l'analyse d'un corpus d'apprenant est en général validé par comparaison avec un corpus de locuteurs natifs de type équivalent (NS/NNS)[17] ou bien avec un autre corpus d'apprenants de langue maternelle différente (NNS/NNS)[18] mais de niveau comparable. Les possibilités d'études contrastives dépendent essentiellement du niveau des étudiants : ce n'est que pour les productions d'élèves de niveau avancé qu'il est possible de trouver un corpus de locuteurs natifs relativement équivalent. Les différents corpus d'apprenants d'une même langue cible peuvent par contre être beaucoup plus facilement mis en parallèle.

L'étude synchronique et/ou diachronique peut prendre différentes directions. Le corpus d'apprenants peut être annoté au niveau des erreurs commises et permettre ainsi l'établissement d'une typologie représentative de la population étudiée. Pour repérer et dénombrer les usages erronés ou corrects, les corpus d'apprenants devront être annotés de manière cohérente et systématique (Granger, 1998). Cela prend beaucoup de temps car les outils informatiques développés pour annoter les corpus de locuteurs natifs ne sont pas totalement adaptés à cette tâche dans la mesure où ils ont essentiellement été conçus dans le but de repérer les formes standard de la langue (Barlow, 2005).

Une deuxième alternative pour l'analyse des corpus d'apprenants est constitué par le comptage « des éléments lexico-morpho-syntaxiques qui sont typiquement soit sous-utilisés soit sur-utilisés par les apprenants (Barlow, 2005, p. 335) ».

Le plus souvent, les différents outils disponibles permettent de caractériser l'interlangue sous forme de listes de fréquence d'emploi des mots, expressions ou structures, fautives ou non fautives, de la langue apprise. Il est également possible de repérer l'ensemble des formes totalement évitées par les apprenants car considérées comme trop éloignées de leur langue maternelle.

Dans tous les cas de figure d'analyse, une interprétation des données collectées doit être proposée notamment par le biais de la comparaison avec un corpus de locuteurs natifs.

Repérer et analyser la fréquence d'apparition des formes fautives et/ou correctes est essentiel si l'on veut comprendre l'interlangue associé à une communauté particulière d'apprenants d'une langue étrangère donnée. Cette compréhension orientera ensuite les pratiques pédagogiques appliquées à l'enseignement de cette langue, les stratégies de classe devant de toute évidence tenir compte des difficultés linguistiques qui sont propres à ce groupe.

6 LE CORPUS ANTILLAIS D'APPRENANTS DE FRANÇAIS

Le corpus antillais d'apprenants de français créé au Centre d'apprentissage des langues (CAL) de l'université des Indes occidentales de Trinité et Tobago est constitué de courtes rédactions, de 100 à 250 mots environ, produits par les étudiants des niveaux débutant à intermédiaire avancé (A1 à B1/B2 dans la grille du CECR) du centre apprenant le français. Ces rédactions

[17] NS = Native Speaker
[18] NNS = Non Native Speaker

ont toutes été écrites durant les deux évaluations semestrielles que fait chaque étudiant du CAL. Elles ont été collectées d'octobre 2007 à avril 2009.

Une fois sa transcription terminée, ce corpus comprendra près de 90 000 mots, correspondant à environ 400 écrits d'étudiants. Un nombre significatif d'apprenants ont contribué quatre écrits ou plus au cours de leurs différents semestres d'étude au CAL. Chaque apprenant présent dans le corpus a donné son accord pour l'inclusion de ses écrits. Cette autorisation est renouvelée chaque semestre de présence au centre.

Bien que le corpus ait été rendu totalement anonyme, l'analyse des erreurs commises exige que quelques informations soient enregistrées afin de pouvoir associer certaines formes fautives à des variables telles que le niveau de compétence en français ou bien encore le sujet traité. Chaque apprenant reçoit un nom-code qui permettra de suivre l'évolution de son interlangue au cours de ses semestres d'étude au CAL.

Les apprenants du Centre d'apprentissage des langues de l'université des Indes occidentales constituent un groupe très homogène. La très grande majorité d'entre eux sont originaires des Antilles anglophones, essentiellement de Trinité et Tobago. Les quelques étudiants non anglophones ou non antillais ont été écartés à dessein afin de ne pas influencer les résultats obtenus. Tous les cours de français du CAL utilisent la même série de manuels d'enseignement: *Breakthrough French 1*, *2* et *3*. Les conditions de production sont particulièrement stables. Il s'agit toujours d'écrits réalisés en classe sans aide extérieure ni utilisation de dictionnaires ou de grammaires de référence. Chaque étudiant ne disposait que de trente minutes pour achever sa rédaction. Les apprenants avaient à traiter de sujets tels que se présenter ou donner son opinion sur des questions de vie quotidienne.

Voici, à titre d'exemple, un extrait d'une rédaction écrite par un(e) étudiant(e) de niveau A1, transcrite telle quelle, c'est-à-dire incluant toutes les erreurs lexicales et syntaxiques commises:

« Bonjour Messieurs-dames ! Je suis Trinidadiene. J'habite en Arima avec mes famille. Je suis professeur et je travaille en un école en Arima. Je ne suis pas mariée. Je suis célibataire. Ma mère et mon père sont marié et habitent avec mon frère et moi. […] J'aime, à manger, le sandwich au jambon et le glace au chocolat et à boire, j'aime un lait-chaud. Je n'aime pas le cinéma et la bière. Merci beacoup. »

A1 (CECR)/Novice Low (ACTFL)/Test 1

L'extrait suivant concerne un(e) apprenant(e) du niveau B1 :

« Il y a plusiers monuments remarquables dans P.O.S. Mon favorit est le château Stollemmeyer – un édifice grand et gris à côte de la Savanne. La Savanne est grande et elle s'appelle "Queen's Park Savannah". Il y a beaucoup d'arbres et on peut jouer au sports et se promener la. Quelques fois les agents de police se promene au cheval dans les rues de la capitale. L'embouteillage n'arret pas les cheveaux ! »

B1 (CECR)/Intermediate Mid (ACTFL)/Test 1

7 ANALYSE DU CORPUS

Ce travail vient de débuter et les résultats ici présentés sont préliminaires et visent principalement à donner un aperçu des possibilités d'analyse des productions d'apprenants que permet la linguistique de corpus.

Les quelques particularités décrites ci-dessous ainsi que leur interprétation sont d'abord le fait de l'intuition personnelle de l'auteur de cet article et devront *in fine* être corroborées par une comparaison avec plusieurs corpus de référence qu'ils soient de L1 ou de L2 comme par exemple :

25. le corpus ICE d'anglais jamaïcain,

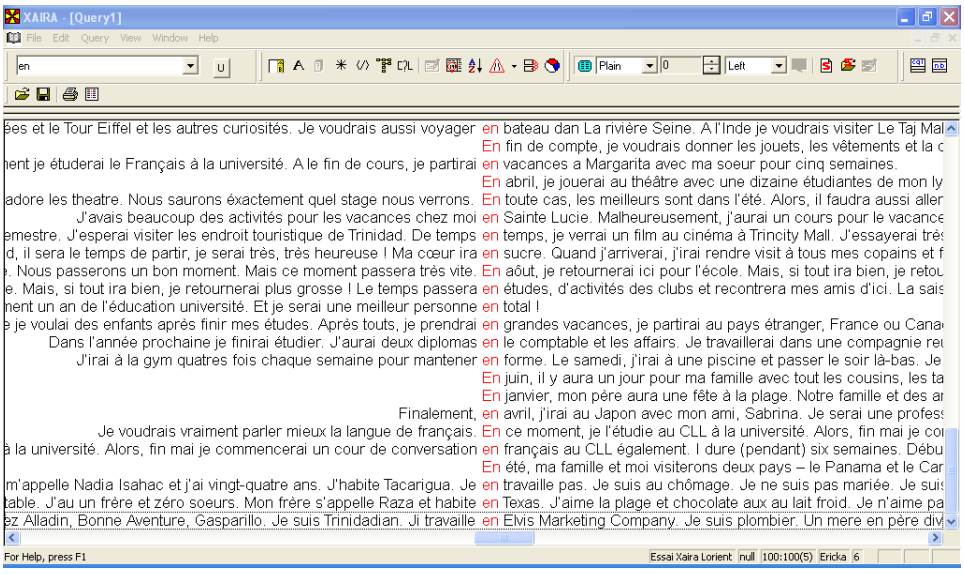
26. un corpus équivalent de français langue maternelle s'il existe,
27. un corpus d'apprenants britanniques de français encore à trouver.

Les analyses présentées ci-dessous ont été réalisées sur le corpus brut, non annoté, à l'aide de l'outil de consultation XAIRA développé par les services informatiques de l'université d'Oxford.

Un premier survol rapide, réalisé lors de la transcription des manuscrits sous forme de documents informatiques, permet de constater le phénomène de la généralisation intralangue : généralisations abusives d'aspects de la langue étudiée. Certaines particularités morpho-syntactiques du français sont ainsi étendues à la production d'autres formes apparentées : ainsi l'apparition d'une graphie telle que « çela » est visiblement influencée par la façon d'écrire le démonstratif « ça » en français. On peut signaler aussi la forme « àu » qui semble résulter de l'association « à + le » enseignée très tôt aux étudiants mais imparfaitement maîtrisée dans les premiers temps de l'apprentissage. Cette généralisation intralangue se manifeste également dans la façon de conjuguer certains verbes notoirement irréguliers, ce qui donne des formes fautives telles que « faisent » et « etez ».

L'intrusion de la langue maternelle (L1) peut expliquer que le genre soit assez peu maîtrisé par les étudiants. Le choix entre féminin et masculin en ce qui concerne les noms paraît bien aléatoire, tout comme les accords en genre et en nombre dans les combinaisons nom/adjectif. Ce problème semble récurrent chez les étudiants dont la langue maternelle, comme l'anglais, ne connaît que partiellement les catégories grammaticales du genre et du nombre. Cette explication vaut également pour les accords sujet/verbe, pour lesquels les désinences verbales sont souvent fautives dans les premiers temps de l'apprentissage pour notre population d'apprenants.

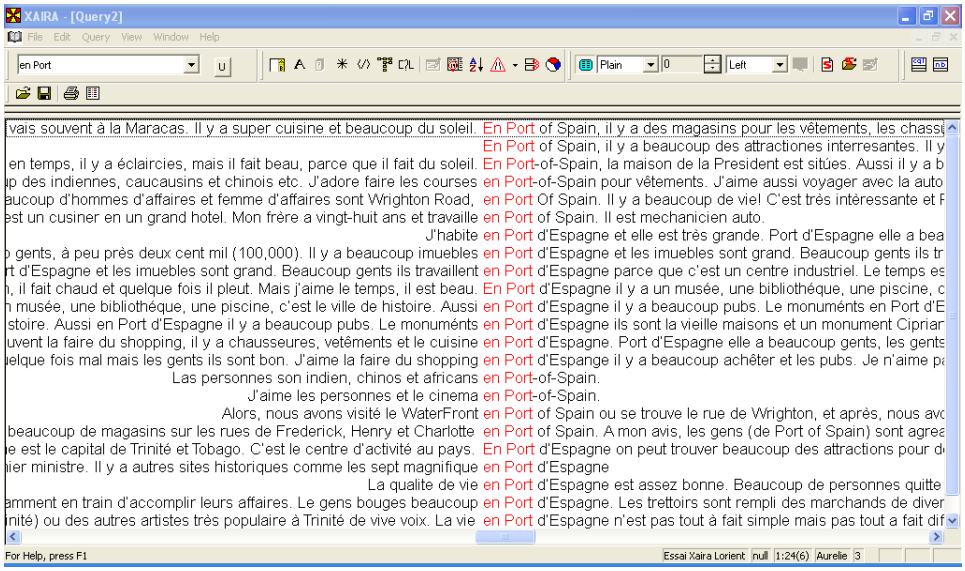
L'influence de la langue maternelle est d'autant plus forte que les ressemblances morpho-syntaxiques entre la L1 et la L2 sont nombreuses. Cette constatation pourrait expliquer la prédominance de la préposition « en » dans les écrits, que les apprenants semble associer automatiquement au « in » anglais en oubliant le « dans » du français.



Copie d'écran 1 : Sur-utilisation fautive de la préposition « en » due à l'influence de l'anglais

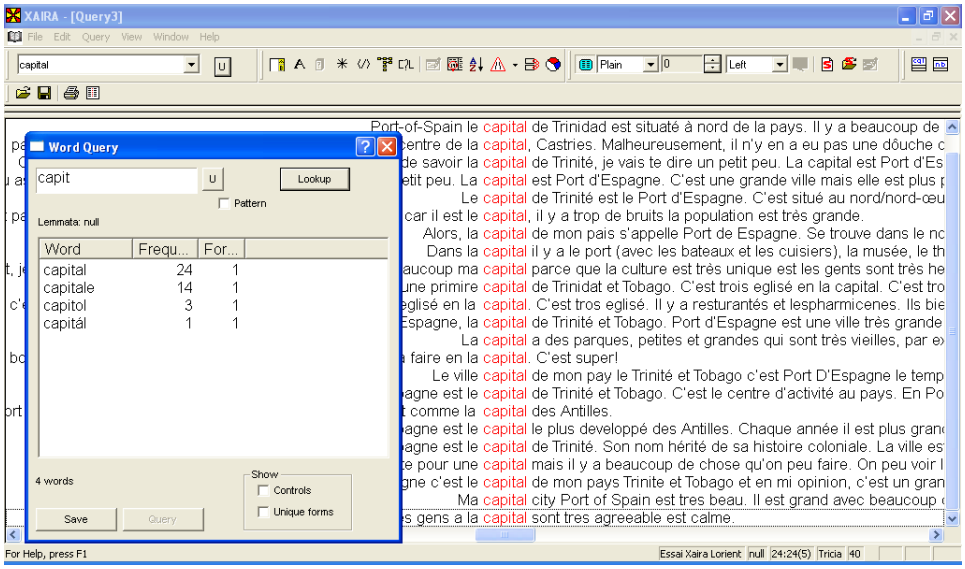
Cette intrusion de la L1 est, à Trinité et Tobago, renforcée par l'influence exercée par l'espagnol, la langue étrangère (une autre L2) dont l'étude est devenue obligatoire dans le

pays. Au niveau syntaxique, cette double influence pourrait expliquer que la préposition française « en » soit presque systématiquement employée notamment dans les contextes linguistiques où en anglais serait utilisée la préposition « in » et dans le même type de phrases la préposition « en » en espagnol. La copie d'écran 1 permet de rendre compte de ce phénomène : les apprenants disent habiter, vivre, aller et travailler « en » Port d'Espagne, capitale de l'état de Trinité et Tobago, et non pas « à » Port d'Espagne.



Copie d'écran 2 : Sur-utilisation fautive de la préposition de lieu « en » due à la double influence de l'anglais et de l'espagnol

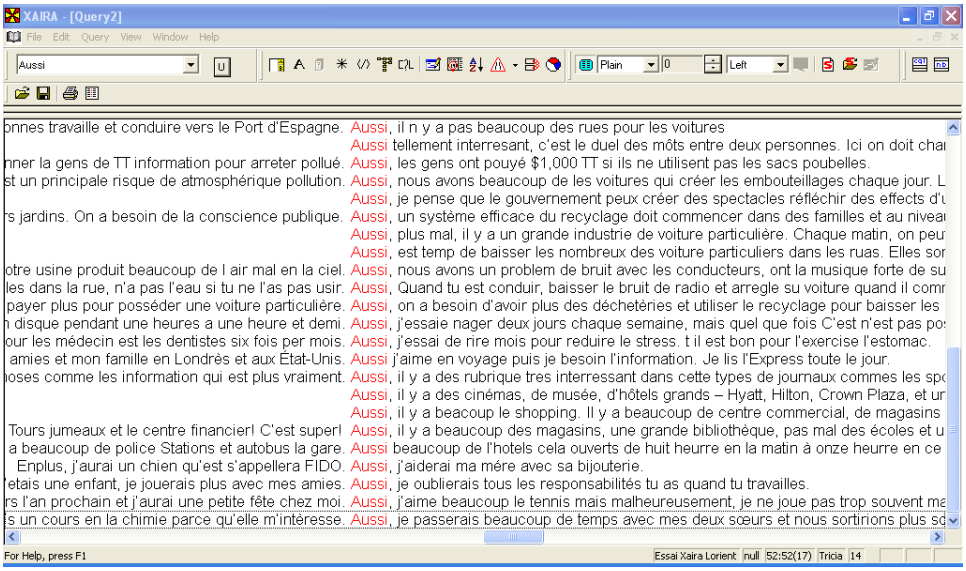
Cette influence croisée se constate également en ce qui concerne l'orthographe. Le mot « capitale » pour prendre une seule occurrence s'écrit, dans la partie du corpus antillais déjà transcrite, majoritairement « capital » comme en anglais et en espagnol.



Copie d'écran 3 : Influence des langues anglaise et espagnole sur la maîtrise de l'orthographe française

L'intrusion de l'anglais n'est peut-être pas toujours renforcée par l'espagnol dans cette population d'apprenants. On peut imaginer en effet que l'influence de l'espagnol puisse contrer, pour certains traits morpho-syntaxiques, celle de la langue maternelle anglaise. Par exemple, les difficultés relevant des catégories du genre et du nombre et des accords qui en découlent, mentionnées ci-dessus, devraient être atténuées pour ces étudiants car ces catégories existent également en espagnol. La remarque est sans doute également valable pour les désinences verbales en français qui comme en espagnol sont omniprésentes.

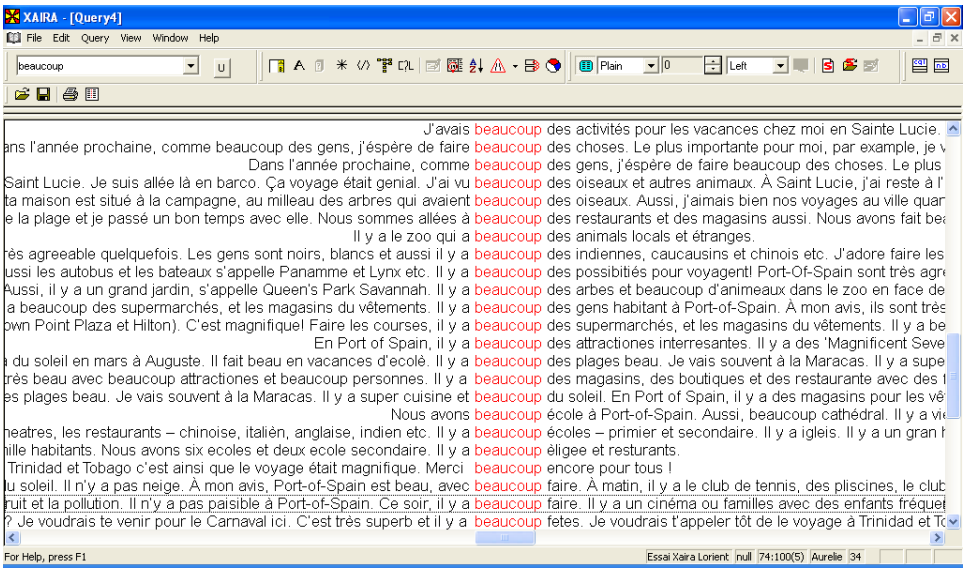
L'intrusion de la seule langue anglaise est sans doute à l'origine de la copie d'écran ci-dessous. Les apprenants du CAL ont tendance à structurer leur paragraphe en commençant certaines phrases par l'adverbe « aussi », ce qui correspond à un ordre syntaxique anglais, inconnu de la langue française.



Copie d'écran 4 : Influence de la langue anglaise présente dans l'utilisation de l'adverbe « aussi » en début de phrase

L'influence des méthodes d'enseignement sur l'interlangue des apprenants, ce « formulaic parroting » décrit par l'ACTFL dans les premiers temps de l'apprentissage, se constate largement dans le corpus. La prise de risque linguistique qui passe par la réutilisation originale des éléments de la langue apprise ne survient généralement qu'aux niveaux plus avancés. Pour décrire leur capitale, Port d'Espagne, les apprenants reprennent ainsi telles quelles les formules présentées dans leur livre comme le « très riche sur le plan du patrimoine ». Ce phénomène se voit aussi dans les associations très fréquentes de mots : le nom « édifice » est ainsi quasi systématiquement suivi de l'adjectif « remarquable ».

Une annotation du corpus au niveau des formes fautives – avec un outil informatique autre que XAIRA – permettrait de faire ressortir une série de difficultés linguistiques rencontrées par cette population particulière d'apprenants comme par exemple l'utilisation des expressions de quantité en français. La copie d'écran 5 ci-après donne une série de collocations qui rendent compte du problème pour « beaucoup de ».



Copie d'écran 5 : Prédominance de certaines difficultés comme l'utilisation de « beaucoup de »

Une fois achevée, cette recherche débouchera sur l'élaboration d'une liste des problèmes lexicaux et morpho-syntaxiques les plus fréquemment rencontrés par les étudiants antillais du CAL dans leur apprentissage du français langue étrangère. Elle devrait également mettre en avant toute une série d'interprétations destinées à rendre compte de l'originalité de l'interlangue produit par cette communauté particulière d'apprenants. Elle permettra enfin une amélioration significative de la pédagogie appliquée à l'enseignement du FLE dans cette région du monde.

8 AVANTAGES POUR LES ENSEIGNANTS

Dans le cadre de sessions de formation continue, les données recueillies par cette recherche seront présentées aux professeurs du CAL. Les enseignants seront également initiés à l'utilisation de l'outil informatique XAIRA afin de faciliter une meilleure appréhension de ces résultats. Si l'idée qu'une langue étrangère ne pose pas forcément les mêmes difficultés d'apprentissage aux diverses populations d'apprenants qui s'y essaient est bien acceptée, il n'est pas facile de pointer du doigt les spécificités de telle ou telle communauté. Il est pourtant essentiel de connaître l'interlangue typique de la population particulière d'apprenants qui vous est confiée face à des méthodes d'enseignement qui le plus souvent n'ont pas été pensées pour celle-ci mais sont en général destinées à des apprenants apatrides, génériques, qui n'existent que dans l'imagination des éditeurs. Les auteurs de méthodes de FLE font bien souvent comme si la culture et l'environnement, par définition très divers, dans lesquels baignent les différentes communautés d'apprenants n'existaient pas ou n'avaient aucune influence sur leur apprentissage de la langue.

Il nous paraît important de pouvoir compléter le livre de textes utilisé en créant du matériel supplémentaire traitant des aspects de la syntaxe, de la morphologie ou du lexique qui nécessitent une attention particulière pour ce groupe d'apprenants de français. Il s'agit d'améliorer l'approche pédagogique en s'adaptant aux difficultés caractéristiques des apprenants trinidadiens en termes de contenus, d'explications ou de progression.

Les enseignants du CAL disposeront enfin de données claires leur permettant d'évaluer les méthodes utilisées en classe et leur propre pratique du français langue étrangère.

9 AVANTAGES POUR LES APPRENANTS

Une pédagogie/didactique d'apprentissage d'une langue étrangère doit impérativement « confronter les apprenants à leurs propres productions » pour « les amener à découvrir par eux-mêmes les traits distinctifs de leur interlangue » (Granger, 2001). Cette approche est parfois appelée éveil au langage ou mise en conscience et s'apparente au « Data-driven Learning » de Johns (1994).

Ce projet de recherche aura pour résultat principal la mise à disposition d'un outil d'investigation du français tel qu'il est écrit par les étudiants antillais trinidadiens de langue anglaise.

Le corpus et ses outils d'analyse seront accessibles aux apprenants, sous forme de CD-Rom ou bien encore en ligne. Cet outil de consultation devra être adapté afin de ne pas rebuter les utilisateurs non initiés à la linguistique informatique. Les apprenants du CAL pourront constater par eux-mêmes le genre de difficultés que leur pose l'apprentissage du français. La pratique régulière d'un tel outil fixera chez les étudiants du centre les aspects linguistiques nécessitant une plus grande attention de leur part. Leur interlangue, élément clé de tout apprentissage langagier (Hanzeli, 1975) devrait alors progresser plus rapidement. Une présentation sur CD-Rom ou en ligne du corpus et des outils de consultation permettra une plus grande diffusion des résultats de la recherche. De plus, le fait d'être toujours disponible devrait également faciliter le travail en autonomie des apprenants.

10 CONCLUSION

Cette recherche, toujours en cours, ne devrait s'achever qu'en fin d'année 2010. Elle a pour principaux objectifs de décrire l'interlangue caractéristique des apprenants trinidadiens de français et de les aider à apprendre cette langue de façon plus ciblée. Elle devrait également contribuer à l'étude sur l'acquisition des langues étrangères bien qu'il se pose un problème de validité car ce corpus antillais d'apprenants de français est de petite taille (Sinclair, 1991).
Les quelques analyses et interprétations initiales des données confirment certaines premières impressions de départ: les apprenants trinidadiens de français n'abordent pas cette langue de la même façon que d'autres communautés de langue anglaise. Leur variété particulière d'anglais, la proximité de l'Amérique latine et une culture aux origines multiethniques expliquent en partie les difficultés et facilités particulières rencontrées par cette population.

11 RÉFÉRENCES

« Cambridge Learner Corpus », dans Cambridge International Corpus, http://ww.cambridge.org/elt/corpus/learner_corpus.htm.

« International Corpus of Learner English-ICLE », dans Centre for English Corpus Linguistics-CECL, http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm.

« The Longman Learners' Corpus », dans Longman Dictionaries, http://ww.pearsonlongman.com/dictionaries/corpus/learners.html.

Barlow M. (2005). « Computer-based analyses of learner language ». Dans R. Ellis et G. Barkhuizen (®ds.), *Analysing Learner Language*. Oxford : Oxford University Press.

Biber D., Conrad S. et Reppen R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge : Cambridge University Press.

Ellis R. et Barkhuizen G. (®ds.) (2005). *Analysing Learner Language*. Oxford : Oxford University Press.

Flowerdew L. et Tong K. K. (®ds.) (1994). *Entering Text. Hong Kong : The Hong Kong University of Science and Technology.*

Gougenheim G., Michea R., Rivenc P. et Sauvageot A. (1956). *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base.* Didier.

Granger S. (2001). « Didactique des langues étrangères, linguistique de corpus et traitement automatique des langues ». Dans M. Marquillo Larruy (éd.), *Questions d'épistémologie en didactique du français (langue maternelle, langue seconde, langue étrangère),* Poitiers : Cahiers FORELL.

Granger S. (éd.) (1998). *Learner English on Computer.* Longman.

Hanzeli V. E. (1975). « Learner's Language: Implications of Recent Research for Foreign Language Instruction ». *The Modern Language Journal,* vol. 59, no. 8, p. 426-432.

Jantunen J. H. « International Corpus of Learner Finnish ». *Corpus Study on Language-Specific and Universal Features in Learner Language*, http://www.oulu.fi/hutk/sutvi/oppijankieli/en/ICLFI_Corpus.html.

Johns T. (1994). « From printout to handout: grammar and vocabulary in the context of data-driven learning ». Dans T. Odlin (éd.), *Perspectives on Pedagogical Grammar.* Cambridge : Cambridge University Press.

Marquillo Larruy M. (éd.) (2001). *Questions d'épistémologie en didactique du français (langue maternelle, langue seconde, langue étrangère).* Poitiers : Cahiers FORELL.

Milton J. et Chowdhury N. (1994). « Tagging the interlanguage of Chinese learners of English ». Dans L. Flowerdew et K. Tong (éds.), *Entering Text*. *Hong Kong : The Hong Kong University of Science and Technology.*

Nelson G. *International Corpus of English,* http://www.ucl.ac.uk/english-usage/ice/.

Odlin T. (1989). *Language Transfer.* Cambridge : Cambridge University Press.

Odlin T. (éd.). (1994). *Perspectives on Pedagogical Grammar.* Cambridge : Cambridge University Press.

Partington A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching.* Philadelphia : John Benjamins.

Saussure de F. (1964). *Cours de Linguistique Générale.* Paris : Payot.

Sinclair J. (1991). *Corpus, Concordance, Collocation.* Oxford : Oxford University Press.

Sinclair J. (2004). *How to Use Corpora in Language Teaching.* Philadelphia : John Benjamins.

LE VIEILLISSEMENT NORMAL ET PATHOLOGIQUE DU LANGAGE : ÉTUDE COMPARATIVE DES DISCOURS ORAUX

Hye Ran Lee[1], Melissa Barkat-Defradas[1] et Frédérique Gayraud[2]
Laboratoire Praxiling – UMR 5267 CNRS et Université de Montpellier
Laboratoire Dynamique du Langage – UMR 5596 CNRS et Université de Lyon

RÉSUMÉ

La maladie d'Alzheimer est une maladie neurodégénérative, caractérisée par une altération des fonctions cognitives, et donc des fonctions langagières. Nous avons tenté de déterminer des troubles d'ordre syntaxique chez les patients atteints de la maladie d'Alzheimer. Pour ce faire, nous avons mené une étude comparative des discours oraux des patients souffrant de la maladie d'Alzheimer *vs.* patient MCI *vs.* sujets âgés sains *vs.* sujets jeunes. Nous avons ensuite analysé la complexité syntaxique dans un corpus oral recueilli auprès de ces différentes populations à l'aide du logiciel Hyperbase. Les résultats montrent que la complexité syntaxique diminue avec l'âge, et plus fortement en cas de pathologie démentielle. Bien que la production des patients déments soit grammaticalement bien formée, une réduction de la complexité syntaxique et une diminution de la diversité des structures syntaxiques utilisées sont observées. Cette étude pilote tente de démontrer que la réduction de la complexité syntaxique peut constituer un marqueur précoce de la maladie d'Alzheimer et ouvre des perspectives quant à l'utilisation du traitement automatique du corpus dans une mise en pratique clinique.

1 INTRODUCTION

> « In any well-made machine one is ignorant of the working of most of the parts – the better they work the less we are conscious of them…it is only a fault which draws our attention to the existence of a mechanism at all » (Craik, 1943)[19].

Avec le vieillissement de la population, l'incidence des maladies neurodégénératives, telles que la maladie d'Alzheimer, augmente de façon exponentielle. Aujourd'hui, la diminution progressive des capacités langagières est reconnue comme une importante manifestation clinique de la maladie d'Alzheimer qui apparaît comme l'un des symptômes les plus précoces (Martin, 1987 ; Murdoch et *al*., 1987 ; Bayles et *al*., 1991; Orange, 1991 ; Moreaud et *al*., 2001). Les troubles du langage ne se limitent pas aux manifestations de surface ; ils résultent de déficits sous-jacents. Ainsi, le langage des patients souffrant de la maladie d'Alzheimer se montre prometteur pour un nouvel outil de diagnostic précoce de cette maladie et de mesure du déclin des fonctions cognitives (Bayles, 1982 ; Bayles et *al*., 1987 ; Cardebat et *al*., 1991).

La production langagière est une activité cognitive complexe qui nécessite la gestion presque simultanée et en temps réel de diverses sous-composantes telles que la génération et l'organisation du contenu, la lexicalisation et l'exécution articulatoire qui sollicitent à la fois

[19] « Une machine bien construite permet d'oublier le fonctionnement des parties qui la composent – mieux cela fonctionne, moins l'on est conscient de celles-ci… c'est seulement lorsque survient un incident que notre attention est attirée par l'existence d'un mécanisme sous-jacent», notre traduction.

la mémoire à long terme et la mémoire de travail (Coirier et *al.*, 1996). Ainsi, le discours oral, qui est une forme naturelle de communication, donne des informations précieuses sur l'intégration des capacités cognitivo-linguistiques. L'analyse des phénomènes *on-line* dans le discours, c'est-à-dire la dynamique de la production langagière en temps réel, permettra de relever les indices linguistiques pertinents pour constituer un pattern linguistique spécifique dans la maladie d'Alzheimer qui peut contribuer au diagnostic précoce de la maladie d'Alzheimer.

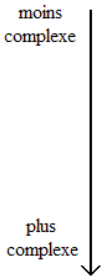
Cependant, malgré la fréquence et l'importance des troubles du langage chez les patients atteints de la maladie d'Alzheimer, les échelles cognitives traditionnellement utilisées pour poser ce diagnostic ne comportent que peu de critères linguistiques et se limitent souvent au niveau lexical (Grossman et *al.*, 1996). La difficulté majeure pour inclure le discours des patients Alzheimer dans un critère de diagnostic est que l'analyse du discours est coûteuse, fastidieuse et demande une certaine connaissance linguistique pour l'interprétation du discours. Le développement de l'analyse du discours automatique permettrait ainsi la prise en compte du discours dans la pratique clinique. Nous allons ici discuter les enjeux et les perspectives des outils d'analyse du discours automatique dans l'étude linguistique sur la maladie d'Alzheimer, à partir de l'analyse comparative des données orales de sujets jeunes *vs*. sujets âgés sains *vs*. populations à risque (patients MCI) *vs*. patients souffrant de démences de type Alzheimer.

Parmi divers aspects linguistiques affectés par la maladie d'Alzheimer, nous allons nous intéresser aux aspects syntaxiques du discours. En effet, du fait que l'apparition de déficits lexico-sémantiques, tels que les fréquents phénomènes anomiques, sont un symptôme spectaculaire dans la maladie d'Alzheimer, ils ont donné lieu à un grand nombre d'études (Appell et *al*., 1982 ; Cummings et *al*., 1985 ; Cardebat et *al*., 1995 ; Barkat-Defradas et *al*., 2008). En revanche, les aspects syntaxiques, c'est-à-dire l'aptitude à manipuler les éléments linguistiques et à les structurer selon les règles de syntaxe pour former des énoncés grammaticalement corrects, apparaissent plus tardivement dans la maladie d'Alzheimer (Kemper et *al.*, 1993; Diaz et *al.*, 2004; Kaprinis et *al.*, 2007), donnant ainsi lieu à peu d'études. C'est pourquoi nous nous intéressons dans notre travail à la détermination des troubles d'ordre syntaxique chez les patients atteints de la maladie d'Alzheimer. Plus précisément, nous tenterons de définir la manière dont la dégradation cognitive influe sur la complexité syntaxique.

La complexité syntaxique est définie à partir des études ontogéniques (i.e. acquisition du langage) et diachronique (i.e. changement historique) : de la construction simple à la construction complexe (i.e. mots < construction simple < construction parataxique (i.e. juxtaposition) < construction syntaxique (i.e. proposition complexe, proposition enchâssée) (Givón, 2009).

Cependant, il n'est pas facile d'articuler la corrélation entre la complexité syntaxique définie par les linguistes et la complexité cognitive définie par les psychologues. Givón (2009) postule une corrélation possible entre la complexité linguistique et la complexité cognitive : les événements représentés mentalement plus complexement sont codés par une structure syntaxique plus complexe ; les événements représentés mentalement plus complexement requièrent une opération mentale plus complexe ; les structures syntaxiques plus complexes requièrent une opération mentale plus complexe. Les propositions enchâssées ou subordonnées, en imposant des exigences supplémentaires pour l'accord entre le sujet et le verbe, les choix pronominaux, la commande linéaire des adjectifs, et d'autres règles grammaticales, augmentent la charge cognitive sur la mémoire de travail. La complexité syntaxique est donc liée au nombre et aux types d'enchâssements dans le discours, ainsi qu'à

leur profondeur. Ainsi, l'échelle de complexité de la construction définie par de nombreuses études va de la construction simple à la construction complexe, en passant par une construction parataxique (O'Donnell, 1974 ; Kroll, 1977 ; Chafe, 1982) :



- Juxtaposition: propositions introduites sans conjonction de coordination ou connecteur.
- Coordination : deux propositions sont coordonnées par l'utilisation d'une conjonction de coordination ou par un autre connecteur.
- Subordination : deux propositions sont combinées, l'une des propositions étant dépendante de l'autre. Le verbe dans les deux clauses est fléchi. Cette classe inclut les complétives introduites par *que*, les propositions introduites par une conjonction de subordination (*parce que*, *lorsque*, *afin que*, *comme*, *si*, *quand*, *où*, *comment*) et les constructions relatives.

Notre hypothèse est donc que les troubles cognitifs se manifestent au plan de la complexité syntaxique et donc que la complexité syntaxique des sujets pathologiques est moins élevée que celle des sujets sains.

2 MÉTHODE

2.1 Sujets

Pour tester notre hypothèse, nous avons effectué une analyse comparative des discours oraux. 20 sujets ont participé à notre étude. 5 patients diagnostiqués comme atteints de démence de type Alzheimer (DTA, 2 femmes et 3 hommes), 5 patients Mild Cognitive Impairment (MCI, 4 femmes et 1 homme), 5 sujets âgés sains (3 femmes et 2 hommes), et 5 sujets jeunes (4 femmes et 1 homme) ont participé. L'âge moyen des patients Alzheimer est de 78, 4 ans, l'âge moyen des patients MCI est de 77, 3 ans, celui des sujets âgés sains est de 79, 3 ans, et enfin la moyenne d'âge des sujets jeunes est de 20, 4 ans. Les 3 groupes de sujets âgés ont un niveau d'étude correspondant au certificat d'étude. Et les sujets jeunes sont en 2ème année de licence. Le score moyen de l'examen neuropsychologique *Mini Mental State Examination* (MMSE, Folstein et *al.*, 1975) est de 22 sur 30 pour les patients Alzheimer, ce qui signifie qu'ils sont au stade léger de la démence, 26 pour les patients MCI, 29 pour les sujets âgés sains, et 30 pour les sujets jeunes. Le tableau 1 récapitule le profil des participants.

	Patients DTA	Patients MCI	Sujets âgés sains	Sujets jeunes
Nombre /Sexe	5 2=F, 3=M	5 4=F, 1=M	5 3=F, 2=M	5 4=F, 1=M
Âge moyen	78,4	77,3	79,3	20,4
Niveau d'étude	Certificat d'étude	Certificat d'étude	Certificat d'étude	Bac +2
Score moyen de MMSE (score max=30)	22	26	29	30

Tableau 1 : Profils des sujets

2.2 Procédure

Nous avons collecté les discours oraux de ces sujets lors d'un entretien individuel. Cet entretien consiste à relever les informations personnelles des sujets, la passation de l'examen neuropsychologique MMSE, et nous avons sollicité la production des sujets à l'aide de deux

tests de mémoire autobiographique : Memory Characteristics Questionnaire (MCQ, Johnson et *al.*, 1988), qui permet de susciter le discours sur l'expérience subjective lors d'évènements spécifiques à l'aide d'indice émotionnel et Autobiographical Memory Test (AMT, Williams et *al.*, 1986), qui permet de susciter le discours sur des souvenirs spécifiques rappelés en réponse à des indices lexicaux.

Tous les entretiens ont été enregistrés et transcrits manuellement. Ainsi, notre corpus est constitué de 20 textes de transcription des discours oraux. Ces transcriptions n'ont pas été analysées individuellement, mais collectivement selon le groupe de sujet.

Une fois le corpus établi, nous l'avons analysé à l'aide d'un logiciel de « traitement documentaire et statistique des corpus textuels », à savoir Hyperbase version 8.0 (CNRS-UNSA) avec lemmatisation par le logiciel d'étiquetage Cordial version 14.0 (Synapse développement). Cette association offre de nombreuses possibilités d'exploration automatique du corpus.

3 ANALYSE ET RÉSULTAT

3.1 Analyse des constructions subordonnées

Nous avons effectué d'abord une analyse sur les propositions subordonnées. Hyperbase a relevé toutes les propositions subordonnées, étiquetées par le logiciel Cordial, et a constitué une liste de distribution de fréquence. L'histogramme produit à partir du résultat (voir figure 1) permet de dresser le profil caractéristique de chaque texte. Ce résultat semble soutenir notre hypothèse : les groupes de sujets pathologiques (patients DTA et patients MCI) ont produit un nombre de propositions subordonnées proportionnellement moins important que les groupes de sujets sains (sujets âgés sains et sujets jeunes). C'est le groupe de sujets jeunes qui a produit la complexité syntaxique la plus élevée, puis le groupe de sujets âgés sains occupe la deuxième place. La complexité syntaxique semble touchée par le vieillissement. Cependant, ce résultat montre également une opposition claire entre le vieillissement normal et pathologique. En effet, la complexité syntaxique du groupe de patients atteints de la maladie d'Alzheimer est beaucoup plus basse par rapport au groupe de personnes âgées saines.

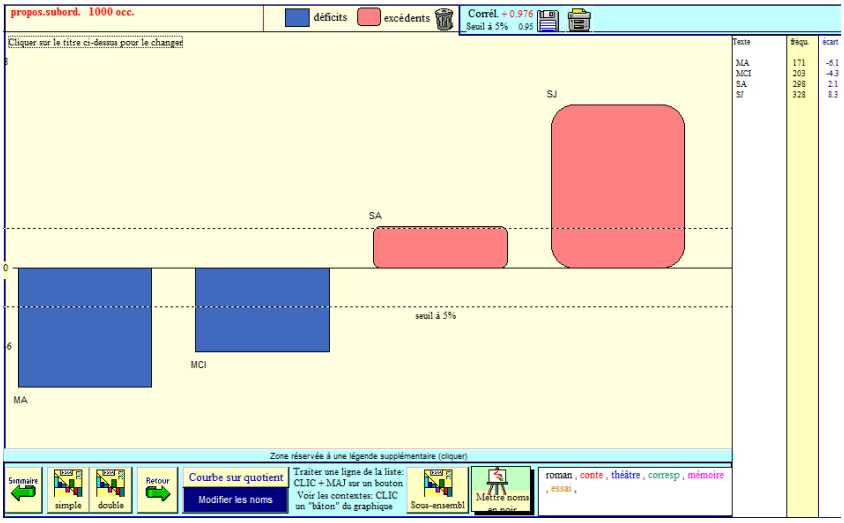


Figure 1 : Distribution de propositions subordonnées selon les groupes[20]

[20] MA (patients atteint de la maladie d'Alzheimer), MCI (patients Mild Cognitive Impairment), SA (sujets âgés sains), SJ (sujets jeunes).

3.2 analyse sur toutes les propositions et les phrases dans le corpus

Pour connaître la tendance et le degré de sophistication du discours de chaque groupe de sujets, nous avons fait sous Hyperbase une requête pour relever toutes les formes de propositions et de phrases étiquetées dans le corpus. L'analyse factorielle produite à partir des résultats (voir figure 2) montre clairement la distance des 4 groupes selon la prévalence des différentes constructions syntaxiques. Le groupe des patients Alzheimer s'oppose aux deux autres groupes de sujets âgés (MCI et âgés sains) ; le groupe des patients MCI et le groupe des sujets âgés sains se rapprochent ; enfin, le groupe des sujets jeunes présente une distance par rapport aux 3 autres groupes.

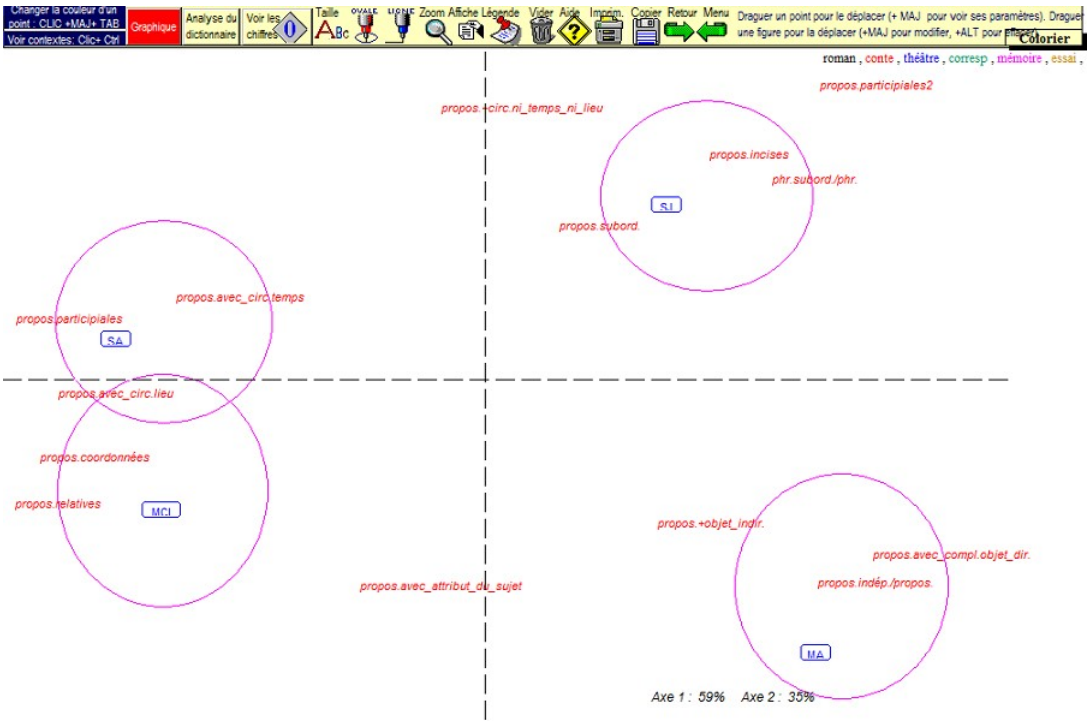


Figure 2 : Résultat de l'analyse factorielle sur toutes propositions et phrases dans le corpus

La spécificité de la structure syntaxique du groupe des patients Alzheimer par rapport aux autres groupes est l'utilisation plus élevée de propositions indépendantes, de propositions avec complément d'objet directe, et de propositions avec objet indirect. Ainsi, les patients Alzheimer ont tendance à s'exprimer en utilisant une structure syntaxique simple par rapport aux autres groupes. Le groupe de patients MCI a été caractérisé par les propositions coordonnées et les propositions relatives. Ce qui signifie que les patients MCI utilisent une structure syntaxique plus complexe par rapport au groupe de patients Alzheimer. La particularité du groupe de sujets âgés sains est la proportion élevée des propositions avec circonstance de temps et de lieu, et les propositions participiales. Enfin, le groupe de sujets jeunes est caractérisé par l'utilisation de propositions subordonnées, et de propositions incises. Nous avons reproduit dans le tableau 2 le degré de présence et d'absence de ces propositions dans le discours des groupes à partir des résultats.

	Groupe de patients DTA	Groupe de patients MCI	Groupe de sujets âgés sains	Groupes de sujets jeunes
Propositions indépendantes	+++	--	---	+
Propositions + complément objet direct	+	-	-	+
Propositions + objet indirect	+	+	-	+
Propositions coordonnées	--	++	+	--
Propositions relatives	-	+	+	-
Propositions + circonstance temps	--	-	+	+
Propositions + circonstance lieu	-	+	+	-
Propositions participiales	-	+	++	-
Propositions subordonnées	--	-	+	++
Propositions incises	-	-	-	+

Tableau 2 : Présence/absence de types de propositions selon les groupes[21]

4 DISCUSSION

Le résultat montre l'opposition entre les groupes de sujets pathologiques et les groupes de sujets sains. La complexité syntaxique du groupe des patients atteints de la maladie d'Alzheimer est moins élevée par rapport aux autres groupes de sujets, et la diversité de la construction syntaxique dans le discours des patients Alzheimer est faible.

La production langagière syntaxiquement complexe requiert vraisemblablement d'importantes ressources de traitement. Cette étude préliminaire a permis de constater que la syntaxe se simplifie dans le discours des patients Alzheimer, certaines formes syntaxiques sont délaissées car trop lourdes cognitivement, comme les propositions participiales et les propositions incises qui demandent au locuteur d'anticiper, de planifier, de garder en mémoire la phrase partiellement construite en même temps qu'il traite la proposition enchâssée.

Nous avons observé également une distance plus étroite entre le groupe des patients MCI et le groupe de sujets âgés sains qu'entre le groupe de patients MCI et le groupe de patients DTA selon leurs choix de construction syntaxique. Ce qui confirme les études précédentes sur l'apparition tardive des perturbations syntaxiques. Cependant, nous avons observé une simplification syntaxique chez les patients atteints de la maladie d'Alzheimer avant même l'apparition de réels troubles syntaxiques entraînant des erreurs grammaticales et des erreurs dans la construction de la phrase.

[21] Un + correspond à un écart réduit de 0 à 5 (inversement pour le -).

5 CONCLUSION

Sans donner de conclusion hâtive, on peut cependant souligner que la complexité syntaxique semble diminuer avec le vieillissement, et plus fortement en cas de pathologie démentielle. Ce résultat montre la possibilité de différenciation entre les sujets âgés sains et les patients atteints de la maladie d'Alzheimer à travers l'analyse des productions langagières.

L'apport de l'analyse automatique dans cette étude est d'abord un gain de temps. En effet, l'analyse manuelle longue, fastidieuse et parfois subjective fait souvent obstacle à la prise en compte des phénomènes linguistiques dans l'étude de la maladie d'Alzheimer. L'utilisation des outils automatiques du discours dans cette étude a montré l'avantage de la détection rapide de critères linguistiques discriminants pertinents et une perspective de mise en pratique clinique de ces critères.

Enfin, des ensembles de données plus volumineux et des variables mieux contrôlées, permettraient de relever un éventail de critères robustes pour le diagnostic précoce de la maladie d'Alzheimer sur la base d'indices linguistiques fiables. Ce qui pourrait, d'un point de vue applicatif, sans aucun doute contribuer à élargir la gamme des diagnostics officiels de la maladie d'Alzheimer, à préciser la sémiologie en vue d'une prise en charge thérapeutique.

6 RÉFÉRENCE

Appell J., Kertesz A. et Fisman M. (1982). « A study of language function in Alzheimer patients ». *Brain and Language,* 17, p. 73-91.

Barkat-Defradas M., Martin S., Rico Duarte L. et Brouillet D. (2008). « Les troubles de la parole dans la maladie d'Alzheimer ». *26ème Journées d'Études de la parole*. Avignon.

Bayles K. A. (1982). « Language function in senile dementia ». *Brain and Language,* 16, p. 265-280.

Bayles K. A. et Tomoeda C. K. (1991). « Caregiver report of prevalence and appearance order of linguistic symptoms in Alzheimer's patients ». *The Gerontologist*, 31, p. 210-216.

Bayles K. et Kazniak A. (1987). *Communication and cognition in normal aging and dementia.* Boston : College Hill Press.

Cardebat D., Demonet J. F., Puel M., Nespoulous J. L., et Rascol A. (1991). « Langage et démences ». Dans M. Habib, Y. Janette et M. Puel (éds.). *Démences et syndromes démentiels : approche neuropsychologique*. Paris : Masson. p. 153-164.

Chafe W. (1982). « Integration and involvement in speaking, writing and oral literature ». Dans D. Tannen (éds.), *Spoken and written language: Exploring orality and literacy,* 35–53, Norwood, New Jersey: Ablex.

Coirier P., Gaonac'h D. et Passerault J. M. (1996). *Psycholinguistique textuelle. Approche cognitive de la compréhension et de la production de texte.* Paris : Armand Colin.

Craik J. K. (1943). *The Nature of Explanation.* Cambridge : Cambridge University Press.

Cummings J. L., Benson D. F., Hill M. A. et Read S. (1985). « Aphasia in dementia of the Alzheimer type ». *Neurology*. 35, p. 394-397.

Diaz M., Sailor K., Cheung D. et Kuslansky G. (2004). « Category size effects in semantic and letter fluency in Alzheimer's patients ». *Brain and language*, 89, p. 108-114.

Folstein M. F., Folstein S. E. et McHugh P. R. (1975). « Mini-mental state. A practical method for grading the cognitive state of patients for the clinician ». *J Psychiatr Res,*12, p. 189-98.

Givón T. (2009). « Introduction. ». Dans. T. Givón et M. Shibatani (éds.), *Syntactic Complexity*. Diachrony, acquisition, neuro-cognition, evolution. Amsterdam/ Philadelphia : John Benjamins Publishing Company. p. 1-19.

Grossman M., D'Esposito M., Hughes E., Onishi K., Biassou N., White-Devine T. et Robinson K. M. (1996). « Language comprehension difficulty in Alzheimer's disease, vascular dementia, and fronto-temporal degeneration ». *Neurology,* 47, p. 183-189.

Johnson M.K., Foley M., Suengas A. et Raye C. (1988). « Phenomenal characteristics of memories for perceived and imagined autobiographical events ». *Journal of Experimental Psychology: General, 117*, p. 371-376.

Kaprinis S. et Stavrakaki S. (2007). « Morphological and syntactic abilities in patients with Alzheimer's disease ». *Brain and language*, 103 (1-2), p. 59-60.

Kemper S., LaBarge E., Ferraro R., Cheung H. et Storandt M. (1993). « On the preservation of syntax in Alzheimer's disease ». *Archives of Neurology*, 50, p. 81-86.

Kroll B. (1977). « Combining ideas in written and spoken English: A look at subordination and coordination. ». Dans E. Keenan et T. Bennett (éds.), *Discourse across time and space*, Los Angeles : University of Southern California: Southern California Occasional Papers in Linguistics. p. 69-108.

Martin A. (1987). « Representation of semantic and spatial knowledge in Alzheimer's patients : Implications for models of preserved learning and amnesia ». *Journal of Clinical and Experimental Neuropsychology*, 9, p. 121-124.

Moreaud O., David D., Charnallet A. et Pellat J. (2001). « Are semantic errors actually semantic? Evidence from Alzheimer's Disease ». *Brain and Language*, 77, p. 176-186.

Murdoch B. E., Chenery H. J., Wilks V. et Boyle R. S. (1987). « Language disorders in dementia of the Alzheimer type ». *Brain and Language,* 31, p. 122-137.

O'Donnell R. C. (1974). « Syntactic difference between speech and writing ». *American Speech* 49, p. 102–110

Orange J. B. (1991). « Perspectives of family members regarding communication changes ». Dans R. Lubinski (éd.), *Dementia and communication*. Philadelphia, PA : B.C. Decker. p. 98-114.

Petersen R. C., Smith G. E., Waring S. C., Ivnik R. J., Tangalos E. G.et Kokmen E. (1999). « Mild cognitive impairment: clinical characterization and outcome ». *Arch Neurol*, 56, p. 303-308.

Williams J. M. et Broadbent K. (1986). « Autobiographical memory in suicide attempters ». *Journal of Abnormal Psychology, 95*(2), p. 144-149.

UNE ÉTUDE DE CORPUS POUR LA DÉTECTION AUTOMATIQUE DE THÈMES

Laurence Longo et Amalia Todiraşcu
LiLPa – Université de Strasbourg

RÉSUMÉ

Nous présentons un outil de détection automatique de thèmes, paramétrable selon le genre textuel des documents à traiter (rapports, articles de journal, fiches produits). Cet outil sera utilisé pour optimiser l'indexation et la recherche des documents dans une archive de documents internes à une organisation. Le système hybride de détection de thèmes que nous mettons en place combine des méthodes statistiques et linguistiques pour dégager les thèmes de chaque document. En particulier, pour la détection de thèmes, nous exploitons une catégorie spécifique de marqueurs de cohérence : les chaînes de référence. Dans cet article, nous présentons l'étude d'un corpus composé de divers genres textuels permettant de construire les ressources linguistiques nécessaires à l'identification automatique de ces chaînes de référence.

1 CONTEXTE ET MOTIVATION

Pour faire face à la croissance exponentielle du nombre de documents disponibles sur Internet, la plupart des systèmes de recherche d'information en viennent à se tourner vers les techniques de Traitement Automatique des Langues (TAL) qui exploitent les informations syntaxiques ou sémantiques, dans le but d'améliorer la qualité des résultats fournis par les moteurs de recherche (Intuition ; Illouz et *al*, 2000). Parmi le flot de résultats renvoyés à l'issue d'une recherche, rares sont les résultats qui comportent les informations attendues et paradoxalement, certains documents pertinents ne sont pas retrouvés par les moteurs de recherche. Ce manque de pertinence trouve son explication dans la méthode d'indexation par mots-clés utilisée par les moteurs de recherche, qui ne tient pas compte des propriétés linguistiques des textes (syntaxe, sens, genre textuel, etc.).

Nous proposons une méthode d'indexation des documents basée sur l'indexation par thèmes. Bien que n'ayant pas de définition unanime, le thème d'un document est considéré en général comme le sujet (« de quoi il s'agit dans un document ») d'une narration, d'un texte explicatif ou d'une conversation. Dans la littérature, le thème est considéré selon une perspective phrastique (Halliday, 1985), textuelle (Firbas, 1992 ; Daneš, 1974), ou selon le contenu (ou l'expression linguistique explicite) (Lambrecht, 1994). Disséminés au sein de structures discursives complexes, les thèmes sont réalisés par des expressions linguistiques variées. L'humain avisé est capable d'identifier chacune des composantes thématiques permettant de reconstruire les « thèmes composites » (Bilhaut, 2006), mais il n'en est pas de même pour un système informatique, vu la variété des paramètres à prendre en compte pour identifier les thèmes.

Notre moteur de recherche doit proposer à l'utilisateur de notre outil des recherches ciblées sur les acteurs (« Qui ? ») ou les idées (« Quoi ? »), par la mise en relation des termes de sa requête avec les thèmes des documents détectés automatiquement, comme par exemple :

- *Barack Obama … l'élection présidentielle*
- *La satisfaction des clients…des études de marché*
- *Poussette Red Castle … multiplaces … tout terrain*

Ainsi, pour notre projet, les thèmes phrastiques correspondent aux acteurs et aux idées développées. Les thèmes textuels constituent les sujets d'un texte, ou d'un fragment de

document et sont posés comme agrégats des thèmes phrastiques (Goutsos, 1997)[1]. Le module de (Goutsos, 1997) est axé sur la réalisation des thèmes dans le texte. Il établit des relations entre les unités du discours (des expressions linguistiques explicites et des caractéristiques textuelles spécifiques au genre expositif), permettant de détecter la continuité ou la discontinuité thématique dans le texte.

Les études linguistiques proposant des modèles de structuration et d'interprétation du discours s'appuient sur des paramètres comme la cohésion et la cohérence pour identifier les changements thématiques. Définir la cohérence et la cohésion d'un texte demeure assez difficile. D'une part, la cohérence d'un texte se traduit par des règles de « bonne formation du texte » (Charolles, 1994, 1997) : la répétition (utilisation des procédés de rappel de certains éléments d'information), la progression (utilisation des connecteurs, ou des types de progression thématique : les thèmes traités par des paragraphes voisins sont proches, le thème du paragraphe suivant est déjà introduit dans le paragraphe précédent), la non-contradiction (aucun élément sémantique ne doit contredire un contenu posé ou présupposé). D'autre part, la cohésion du texte se traduit par un système de marques assurant une cohésion anaphorique (chaînes de référence (Corblin, 1995)) et thématique : les liens thème-rhème qui se retrouvent dans l'ensemble du texte (Hernandez, 2004). Plusieurs modèles linguistiques proposés par la théorie du centrage (Grosz, et al., 1995), par la théorie de l'accessibilité (Ariel, 1990) ou par la théorie du Donné (Gundel, Hedberg et Zacharski 1993, 2000) caractérisent la cohérence locale du discours en faisant appel à des éléments comme le choix des expressions référentielles (Schnedecker 1997, 2006), divers marqueurs d'accessibilité (Ariel, 2001) ou des marqueurs de continuité ou de rupture thématique (Cornish, 2000).

Des travaux de recherche sur la détection automatique de thèmes s'appuient sur une identification automatique des chaînes lexicales (termes similaires ou synonymes qui se répètent dans le texte) basée sur des méthodes statistiques et des algorithmes de classification (Stokes et al. 2004 ; Sitbon et Bellot, 2005 ; Choi et al., 2001 ; Hearst, 1994), sur l'utilisation de dictionnaires de synonymes (Hirst, 1998), sur l'exploitation d'expressions linguistiques porteuses de thèmes (Ferret et al, 1998 ; Ferret, 2006 ; Pierard et Bestgen, 2006 ; Allen et al., 1998). Les thèmes et les rhèmes sont identifiés automatiquement parmi les groupes nominaux (simples, complexes ou noms propres) à l'aide des marqueurs linguistiques ou extralinguistiques (l'ordre dans la phrase, la fonction syntaxique (sujet, objet, etc.)). En revanche, les liens anaphoriques ou de co-référence (Kleiber, 1994) sont peu exploités pour la détection automatique des thèmes, malgré leur contribution à l'organisation textuelle (Hernandez, 2004). Pour identifier les thèmes, nous exploitons ces marqueurs linguistiques de la cohérence mais aussi des marqueurs de la cohésion et des propriétés spécifiques des genres textuels (Biber, 1994). Les chaînes de référence, *séquences d'expressions singulières apparaissant dans un contexte telle que si l'une des ces expressions réfère à quelque chose, toutes les autres y réfèrent aussi* » [(Chastain, 1975) traduit par (Corblin, 1995)], participent à la cohérence textuelle et constituent un indice fiable pour révéler les thèmes (Victorri, 1999). Nous utilisons les spécificités des chaînes de référence, liées au genre textuel, car nous partageons l'idée de (Lin et Hovy, 1997) selon laquelle chaque genre de texte comporte des régularités spécifiques dans sa structure du discours. Dès lors, nous étudions les chaînes de référence selon des propriétés dépendantes du genre, comme leur longueur, la nature ou la fréquence de leurs maillons. Nous utilisons un corpus composé d'extraits de genres textuels différents dans l'objectif de proposer une typologie de ces chaînes de référence.

Intégrée à notre architecture de détection des thèmes, cette étude en corpus va nous permettre d'établir une sorte de « profil-type » des chaînes de références issues d'un genre

[1] Nous préférons utiliser la terminologie de (Goutsos, 1997) qui rentre en adéquation avec la procédure de décision mise en place pour détecter les thèmes dans le cadre de notre projet.

textuel particulier. Ces profils vont permettre de paramétrer notre outil pour identifier plus facilement les chaînes de référence issues de chaque genre textuel. Par exemple, notre système sera en mesure de « prédire » le nombre de maillons des chaînes de référence attendu pour un genre particulier, ou bien d'attribuer la priorité (un score élevé) à un candidat appartenant à la catégorie grammaticale « préférée » d'un genre textuel défini.

La suite de l'article est organisée comme suit : nous présentons l'architecture de notre système de détection automatique de thèmes (section 2) puis l'étude des chaînes de référence que nous avons effectuée sur un corpus issu de genres textuels variés (section **¡Error! No se encuentra el origen de la referencia.**).

2 ARCHITECTURE DU SYSTÈME DE DÉTECTION AUTOMATIQUE DE THÈMES

2.1 Présentation du système

L'outil de détection de thèmes que nous mettons en place permettra d'effectuer, en parallèle à la recherche plein texte classique des moteurs de recherche, une recherche par thèmes. Les documents seront indexés par les thèmes qu'ils contiennent. Nous appliquons une méthode hybride qui combine techniques statistiques et marqueurs linguistiques pour identifier les thèmes (voir ***Figure 1***).).

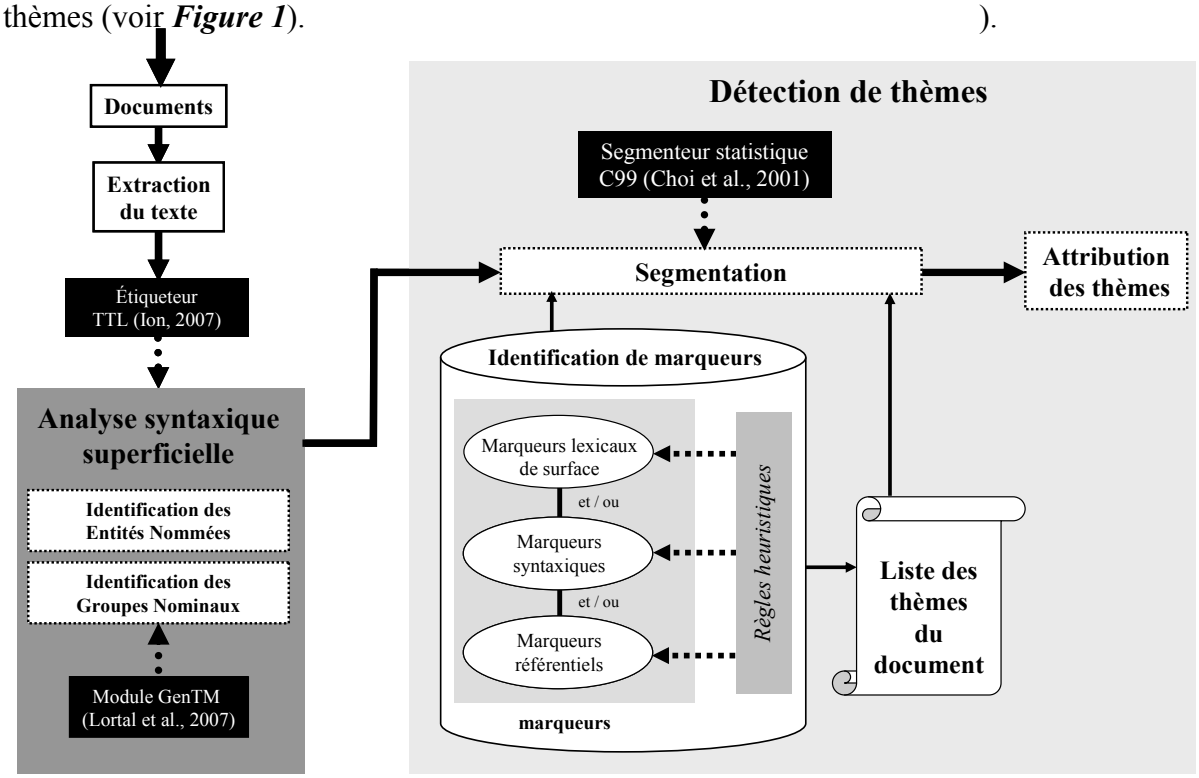


Figure 1 : Architecture du système de détection automatique de thèmes

Ainsi, le texte issu de documents de formats divers (PDF, Office ou HTML) est tout d'abord extrait, puis le texte est segmenté en « segments thématiquement homogènes » (que l'on peut grossièrement rapprocher du paragraphe) par C99[2] (Choi et *al.*, 2001)[3]. Cet algorithme statistique à base de cohésion lexicale utilise une mesure de similarité entre chaque

[2] L'algorithme est disponible à : http://www.freddychoi.co.uk/ (la version 1.3 de C99 bénéficie de l'amélioration de l'Analyse Sémantique Latente (LSA) présentée dans (Choi et *al*, 2001)).

[3] Des méthodes de segmentation statistiques automatiques disponibles, notre choix s'est porté sur C99 car cet outil, de nombreuses fois évalué, s'avère être le plus performant.

unité textuelle, représentée comme vecteur de mots. Bien que C99 fournisse un découpage des textes en segments thématiquement homogènes, en fonction de la distribution et des fréquences de mots, il n'attribue pas de thème à chaque segment. Afin de compléter ce découpage statistique et d'identifier les thèmes de chaque segment, nous appliquons des outils de TAL permettant de dégager des candidats-thèmes.

Pour sélectionner les thèmes associés à chaque segment, nous combinons plusieurs catégories de marqueurs (marqueurs lexicaux de surface, marqueurs syntaxiques, marqueurs référentiels), séparément ou simultanément. Pour l'identification de ces marqueurs et des candidats-thèmes, le texte est étiqueté (Ion, 2007), analysé syntaxiquement (Lortal et *al*., 2007) et annoté au niveau des Entités Nommées (typage des noms de lieu, personne, organisation et fonction) (voir section 3.2). Nous utilisons ces annotations linguistiques pour identifier les anaphores et les chaînes de référence. Le module d'identification des chaînes de référence est basé sur l'échelle d'Accessibilité d'(Ariel, 1990) et combine des informations morpho-syntaxiques (genre, nombre), la position dans la proposition (thème, rhème), ainsi que des propriétés spécifiques liées au genre textuel (voir section 3.3).

Après l'extraction des chaînes de référence de chaque segment, nous considérons comme candidat-thème le premier maillon de chacune des chaînes de référence. Notre but étant de sélectionner des thèmes globaux associés aux documents, nous choisissons parmi les candidats-thèmes ceux répondant à une sélection de critères tels la fréquence (plusieurs chaînes de référence référant à la même entité), la position dans le document, la continuité thématique entre plusieurs segments textuels. La sortie de ce système est une liste de thèmes qui indexent chaque document.

Nous présentons dans la section suivante les divers types de marqueurs linguistiques sélectionnés dans notre système pour identifier les thèmes.

2.2 Marqueurs linguistiques sélectionnés

Les marqueurs d'organisation textuelle de cohésion et de cohérence sont des instructions données au lecteur sur la structure des documents. Pour marquer un changement de thème, les auteurs utilisent des expressions plus explicites que nécessaire (Vonk et al., 1992). De même que (Hernandez, 2004 ; Minel et al., 2001 ; Piérard et Bestgen, 2006), nous pensons que l'efficacité des marqueurs de cohésion et de cohérence s'accroît lorsqu'ils sont employés de manière simultanée. En cela, notre système combine plusieurs catégories de marqueurs : marqueurs lexicaux de surface, marqueurs syntaxiques, marqueurs référentiels (anaphores et chaînes de référence) pour révéler la structure thématique des documents.

- **Les marqueurs lexicaux de surface** (MLS) englobent les champs thématiques (e.g. « concernant X »), les espaces de discours[4] (e.g. « d'une part/d'autre part »), et les domaines qualitatifs (e.g. « en dépit de X »), selon la terminologie de (Charolles, 1997). Nous avons choisi d'identifier ces trois types de cadre de discours car leur fonction consiste à introduire le thème du segment ou, du moins, à apporter des précisions « relatives à ce thème ». Des cadres de discours thématique, nous traitons ceux que (Porhiel, 2005) nomme « marqueurs de thématisation » -unités linguistiques détachées en tête de phrase-, composés d'un introducteur de cadre thématique (Charolles, 1997) et d'un complément souvent nominal. Ces unités linguistiques spécifient explicitement que le complément introduit constitue le thème du segment. Employés en série, les introducteurs thématiques contribuent à la structuration du discours en établissant des liens avec l'amont (relations anaphoriques) et avec l'aval (relations cadratives) (Porhiel, 2005). Par exemple, (Porhiel, 2001) indique que l'introducteur « à propos de X » a un double fonctionnement syntaxique : il peut indexer une ou plusieurs

[4] Notons que (Charolles, 1997), inclut les marqueurs d'intégration linéaire (Turco et Coltier, 1988) dans les espaces de discours.

propositions ou dépendre d'un constituant morphosyntaxique. Sa portée peut donc se limiter à une phrase, mais aussi s'étendre au-delà et gouverner un segment textuel. L'identification automatique de ces MLS nous permet de déduire par exemple que le groupe nominal qui se situe directement à droite de ce MLS est le thème du segment.

- De leur côté, **les marqueurs syntaxiques** constituent les thèmes et rhèmes phrastiques. La position relative et absolue dans la phrase des groupes nominaux et des groupes prépositionnels indique s'il s'agit plutôt d'un thème ou d'un rhème. Ainsi, nous considérons que le thème représente l'ensemble préverbal de la phrase ; le rhème correspondant alors à l'ensemble post-verbal. Par exemple :

M. Léotard réitère *son appui à une candidature de M. Balladur à l'élection présidentielle.*
Thème verbe *Rhème*

Ce deuxième type de marqueur linguistique permet de fournir un candidat-thème phrastique potentiel lorsque sa détection n'a pu être obtenue à l'issue du traitement par les MLS. Dans ce dernier cas, un poids plus fort se verra attribuer à un thème s'il constitue une reprise (totale ou partielle) du rhème le précédant ; comme dans l'énoncé suivant :

*Pour les élections présidentielles, Valéry Giscard d'Estaing prévoit **des changements**. **Ces changements** sont d'ordre politique mais aussi économique.*

Les marqueurs syntaxiques interviennent aussi lors du calcul des chaînes de référence (attribution d'un poids plus fort aux candidats en position sujet pour constituer le premier maillon d'une chaîne de référence).

- Le troisième type de marqueurs, **les marqueurs référentiels,** correspond aux formes de reprises dans le discours ; à savoir les anaphores (Kleiber, 1994) -référence établie entre deux expressions référentielles-, par exemple : *le président… il*, et les chaînes de référence, composées de maillons qui réfèrent au même antécédent. Dans le cadre de notre projet, nous nous intéressons aux relations anaphoriques de coréférence (anaphores nominales et pronominales uniquement)[5].

En suivant (Schnedecker, 1997), nous considérons qu'une chaîne de référence est une relation qui s'établit entre trois expressions référentielles (nommées *maillons*) au moins, par exemple : *Barack Obama… il … le président*. Les noms propres jouent un rôle important dans la structuration du discours car ils participent à la constitution des chaînes de référence. C'est ainsi que les noms propres se trouvent souvent en tête d'une chaîne de référence dans les portraits journalistiques (Schnedecker, 2005). La redénomination d'un nom propre marque une rupture dans la chaîne de référence (Schnedecker, 1997), même s'il s'agit d'un autre point de vue présenté à propos de la personne en question. Lorsqu'une expression référentielle est utilisée (nom propre, pronom, groupe nominal), elle déclenche un « processus de recrutement » particulier du référent exprimé en première mention (Schnedecker, 1997 ; Jenkins, 2002). De ce fait, le groupe nominal démonstratif (e.g. *ce nouveau gouvernement*) renvoie directement au référent sur la base d'un critère de proximité (Kleiber, 1999). De son côté, le pronom « il » instruit de recruter un référent qui soit l'argument d'une proposition saillante : *« pour marquer un fait crucial de cohérence, on va utiliser un pronom pour continuer de parler d'un référent déjà saillant lui-même ou présent dans une situation saillante et l'on va en parler en continuité avec ce qui l'a rendu saillant* » (Kleiber, 1994). Ainsi, par le choix des expressions référentielles présentes dans une chaîne de référence, un indice est donné quant au référent à « garder en mémoire » et qui constitue alors un thème dans cette portion du discours (le *sentential topic* de (Gívon, 1979)). Les chaînes de référence permettront d'identifier un thème de segment lorsqu'aucun MLS ne sera présent dans le texte.

[5] Nous choisissons de ne traiter ni les cas d'anaphores associatives ni les anaphores zéro, ni les cataphores, car leur part dans notre corpus demeure restreinte.

Dans la section suivante, nous présentons en détail une étude des chaînes de référence dans un corpus composé de divers genres textuel. Cette étude va nous permettre de dégager les propriétés spécifiques des chaînes de référence pour un genre donné ; validant notre hypothèse suivant laquelle à un genre textuel correspond un type de chaîne de référence privilégié. Ces propriétés vont être utilisées pour adapter notre outil de détection de thèmes en fonction du genre textuel traité.

3 ÉTUDE EN CORPUS

3.1 Présentation du corpus

Afin d'étudier le lien qui s'établit entre les genres textuels et les chaînes de référence, le corpus utilisé dans cette étude comprend cinq extraits issus de genres textuels différents : journalistique, littéraire, juridique ; répartis de la manière suivante (voir ***Tableau 1***) :

GENRE	SOUS-CORPUS	PERIODE	NOMBRE DE MOTS
Articles de journaux	*Le Monde*	2004	190 015
Editoriaux	*Le Monde Diplomatique*	1980-1988	74 033
roman	*Les trois Mousquetaires (Dumas)*	1844	235 061
lois européennes	*Acquis Communautaire (Steinberger et al.)*	2006	16 701
rapports publics	*la Documentation Française*	2001	36 784
		TOTAL	552 594

Tableau 1 : Répartition du corpus issu de genres textuels différents

Les articles de journaux du *Monde* traitent de la préparation des divers partis politiques à l'élection présidentielle. Une compétition présidentielle est installée et Valéry Giscard D'Estaing réaffirme le principe d'une candidature de l'UDF. De leur côté, les articles du *Monde Diplomatique* abordent la création de deux centres d'étude français -le CERMOC et le CEDEJ- au Proche-Orient. L'extrait issu des *Trois Mousquetaires* illustre l'arrivée de d'Artagnan au Bourg de Meung. Le portrait du jeune homme et de sa monture y sont décrits de manière précise. Dans les lois européennes issues de l'*Acquis Communautaire*, sont abordées les relations à établir entre la Commission Européenne et les autorités des Etats Membres pour que chacune des parties puisse prendre les mesures adéquates en temps voulu. Enfin, les rapports publics de *la Documentation Française* portent sur une comparaison de la satisfaction des usagers des services publics et privés à l'égard des produits commercialisés.

Les sujets des cinq genres textuels étudiés sont divers. Pour pouvoir développer notre système de détection automatique de thèmes, nous étudions les types de chaînes de référence utilisés dans chaque genre de texte, après avoir soumis notre corpus à des traitements : l'étiquetage et la segmentation.

3.2 Etiquetage et segmentation en chunks

Pour l'étiquetage de notre corpus, nous avons utilisé TTL (Ion, 2007) dans sa version française, car il fournit un étiquetage plus fin que la plupart des autres étiqueteurs. En effet, TTL bénéficie du jeu d'étiquettes morphosyntaxiques proposé par le projet MULTEXT (Ide et Véronis, 1994) précisant des informations comme le genre, le nombre, le temps, le mode, la personne. Par exemple, pour la phrase « B. Obama est président » (*Figure 2 : Etiquetage sous TTL pour la phrase : « B. Obama est président »),* TTL nous indique que le mot « est » a pour étiquette « Vaip3s », c'est-à-dire qu'il est conjugué à la troisième personne du présent de l'indicatif et que son lemme est « être ».

MOT	ETIQUETTE	LEMME
B.	Y	B.
Obama	Np	Obama
est	Vaip3s	être
président	Ncms	président

Figure 2 : Etiquetage sous TTL pour la phrase : « B. Obama est président »

En plus de l'étiquetage des noms propres (Np), TTL nous offre une segmentation en *chunks* ou segments non récursifs qui permet d'identifier les groupes nominaux simples et les groupes prépositionnels. Par exemple, TTL propose le découpage en *chunks* suivant pour le groupe nominal « le président » :

```
<w lemma="le"         ana="Da-ms"   chunk="Np#1">        Le </w>
<w lemma="président"  ana="Ncms"    chunk="Np#1">        président </w>
<w lemma="de_le"      ana="Dg-mp"   chunk="Pp#1,Np#2">   des</w>
<w lemma="Etats-Unis" ana="Np"      chunk="Pp#1,Np#2">   Etats-Unis </w>
```

Figure 3 : Découpage en chunk pour le groupe nominal « le président »

Partant du découpage en *chunks* (groupes nominaux et groupes prépositionnels) de TTL, nous souhaitions aussi rendre compte des « groupes nominaux complexes » (GnC), telle la description définie de la ***Figure*** . Un groupe nominal complexe est un groupe nominal modifié par deux groupes prépositionnels au plus ou bien un groupe nominal modifié par une proposition relative. Pour ce faire, nous avons modifié le module GenTM (Lortal et al., 2007) en rajoutant des règles permettant de « regrouper » plusieurs chunks entre eux. A l'issue de ce dernier module, la description définie « Le président des Etats-Unis » est ainsi regroupée en un GnC « GnC#1» :

```
<w lemma="le"         ana="Da-ms"   chunk="Np#1,GnC#1">        Le </w>
<w lemma="président"  ana="Ncms"    chunk="Np#1,GnC#1">        président </w>
<w lemma="de_le"      ana="Dg-fp"   chunk="Pp#1,Np#2,GnC#1">   des </w>
<w lemma="Etats-Unis" ana="Np"      chunk="Pp#1,Np#2,GnC#1">   Etats-Unis </w>
```

Figure 4 : Groupe nominal complexe « le président des Etats-Unis »

En parallèle à cette segmentation, un extracteur d'Entités Nommées (EN) a été mis au point à partir de l'étiquetage fourni par TTL. Il permet d'identifier mais aussi de classifier les EN en quatre catégories : les noms de lieu, de personne, d'organisation, ainsi que les noms de fonction. Par exemple, le nom d'organisation « *lycée Jean Lefèbvre* » est étiqueté par TTL « nom commun (Nc), nom propre (Np), nom propre (Np) ». Un premier patron (Np → Np Np) permet de regrouper les Np « *Jean* » et « *Lefèbvre* » pour n'identifier qu'une seule EN « *Jean Lefèbvre* ». A partir de cette première identification, une règle contextuelle est utilisée pour étiqueter correctement « *lycée Jean Lefèbvre* » en tant qu'organisation (ORG) (voir ***Figure***). Cet extracteur d'EN, non évalué pour le moment, est constitué de 140 règles.

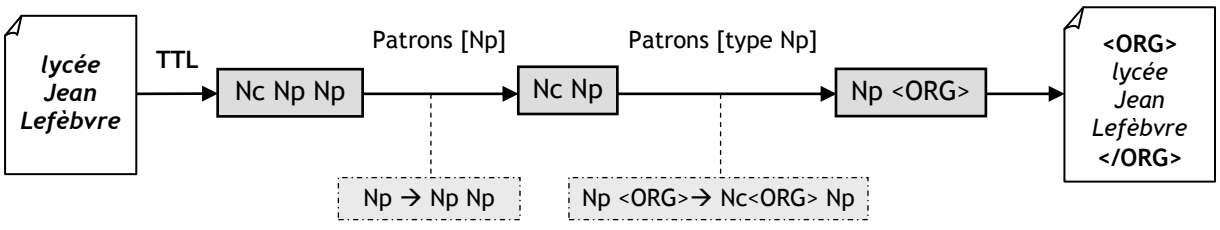


Figure 5 : Identification et typage de l'organisation « lycée Jean Lefebvre »

A l'issue de ces premiers traitements, nous avons pu établir une liste des candidats-thèmes potentiels : Entités Nommées, groupes nominaux simples et complexes (défini, indéfini, démonstratif, possessif), pronoms. A partir de la liste des marqueurs lexicaux de surface et de l'identification des éléments pré et post verbaux (les marqueurs syntaxiques), nous avons

établi une première sélection des candidats-thèmes. Concernant l'identification des marqueurs référentiels, le choix des candidats-thèmes est dicté par la détection des référents. Pour détecter les référents, il faut être en mesure de résoudre la référence, c'est-à-dire de trouver l'ensemble des maillons des chaînes de référence ayant pour référent le thème courant (le principe reste le même pour la détection des antécédents des anaphores). Pour ce faire, nous proposons d'étudier les chaînes de référence dans notre corpus multi genres, car, à l'instar de (Corblin, 1995), nous pensons que les chaînes de référence possèdent des propriétés spécifiques suivant le genre textuel du document.

3.3 Étude des chaînes de référence

La référence et les expressions référentielles (noms propres, descriptions définies / démonstratives, pronoms) ont fait l'objet de nombreuses études, qui se sont focalisées sur leur description linguistique (Kleiber, 1994 ; Cornish, 1995), ou sur leur rôle dans l'organisation du texte (Marandin, 1988 ; Charolles, 1997). Cependant, malgré les nombreux travaux sur la référence, il existe peu de modèles opérationnels pour le français permettant d'identifier automatiquement les chaînes de référence dans les textes (Salmon-Alt, 2001 ; Dupont, 2003 ; Popescu-Belis, 1999).

Dans notre projet, nous mettons en place un module d'identification des chaînes de référence qui aura la particularité d'être paramétrable selon le genre textuel du document. Pour cela, nous étudions les propriétés des chaînes de références sur plusieurs genres textuels. L'étude des chaînes de référence effectuée dans le présent article est inspirée de (Schnedecker, 2005). Ainsi, pour chaque genre, nous examinons les chaînes de référence suivant cinq critères :

- la longueur moyenne (en nombre de maillons) des chaînes de référence,
- la distance moyenne (en nombre de phrases) entre les maillons d'une chaîne de référence,
- la fréquence des maillons des chaînes de référence suivant leur catégorie grammaticale,
- la classe grammaticale des premiers maillons des chaînes de référence ainsi que leur portée,
- la correspondance entre le thème phrastique et le premier maillon des chaînes de référence.

Ces critères vont nous permettre d'établir une sorte de « profil-type » des chaînes de référence issues de chaque genre textuel.

- *La longueur moyenne des chaînes de référence :*

L'étude a révélé quelques différences. Tout d'abord, la longueur moyenne des chaînes de référence varie du simple au triple selon le genre textuel. En effet, les chaînes de référence issues de l'*Acquis Communautaire* comptent en moyenne trois maillons, alors que celles des *Trois Mousquetaires* sont relativement plus longues et comptent neuf maillons en moyenne (voir Tableau 2). Cette différence significative entre la longueur des chaînes de ces deux genres s'explique en ce que les lois européennes font intervenir de nombreux référents, donc que la compétition référentielle est forte. En cas de compétition référentielle, la redénomination d'un nom propre est considérée comme une fermeture du référent en cours (donc une ouverture d'une nouvelle chaîne de référence) (Schnedecker, 2005) ; d'où le faible nombre de maillon pour ce type de chaîne. En revanche, dans l'extrait des *Trois Mousquetaires*, on assiste à de nombreux passages descriptifs, propices aux longues chaînes de référence.

CORPUS	LONGUEUR MOYENNE
Le Monde	4
Le Monde Diplomatique	3,67

Acquis Communautaire	**3**
Les Trois Mousquetaires	**9**
La Documentation Française	3,4

Tableau 2 : Longueur moyenne (en nombre de maillons) des chaînes de référence suivant le genre textuel

- La distance moyenne entre les maillons des chaînes de référence :

L'étude du second critère révèle des disparités quant à la distance moyenne (en nombre de phrases) entre les maillons des chaînes de référence. La distance entre les maillons des chaînes de référence *du Monde* n'excèdent pas une phrase en moyenne, tandis que pour la *Documentation Française*, on assiste à une distance supérieure à deux phrases entre le second maillon et le troisième. Pour ce dernier genre, on retrouve une technique séquentielle particulière : introduction du référent, maintien (ici, on ne trouve pas de compétition référentielle, donc la distance est plus grande entre les maillons 2 et 3) et rappel avant la fermeture de la chaîne de référence (Goutsos, 1997).

CORPUS	RANG DU MAILLON			
	1à 2	2 à 3	3 à 4	4 à 5
Le Monde	0,4	1	1	
Le Monde Diplomatique	0,3	1,3	0	2
Acquis Communautaire	0	1,25		
Les Trois Mousquetaires	0	0,25	0	1,3
La Documentation Française	0,4	**2,4**	0,5	

Tableau 3 : Distance moyenne entre les maillons (en nombre de phrases) suivant le genre textuel[6]

- La fréquence des maillons des chaînes de référence :

Ce critère permet d'identifier la (ou les) catégorie grammaticale privilégiée des maillons des chaînes de référence issues de chaque genre textuel. Ainsi, on constate (***Tableau 4***) une part importante de noms propres (30,8%) dans les maillons des chaînes de référence du *Monde*. Aussi, la moitié des maillons des chaînes de référence du *Monde Diplomatique* sont des groupes nominaux définis (GNdef) alors que 40% des maillons sont des groupes nominaux indéfinis (GNindef) dans l'Acquis Communautaire. Cette dernière observation pour les lois européennes est corrélée au faible nombre de maillons relevés en moyenne dans les chaînes de référence (critère 1). En effet, les mesures décidées par la Commission Européenne ont un caractère générique qui doit s'appliquer à tout Etat Membre de la Communauté ; d'où la présence massive des indéfinis (on aura par exemple : « « un Etat Membre », « une décision », « une mesure »). Les deux derniers genres textuels comptent des pronoms (environ un tiers) et respectivement 28,2% de possessifs (*Les Trois Mousquetaires*) et 33,3% de GN définis (*La Documentation Française*) dans leurs maillons.

CORPUS	CATEGORIE GRAMMATICALE DES MAILLONS					
	Np	Pr	GNdef	GNindef	GNposs	GNdem
Le Monde	**30,8**	15,4	23,1	0	23,1	7,7
Le Monde Diplomatique	0	25	**50**	0	25	0
Acquis Communautaire	0	10	20	**40**	10	20
Les Trois Mousquetaires	0	**35,9**	20,5	10,3	**28,2**	5,1
La Documentation Française	0	**33,3**	**33,3**	16,7	16,7	0

Tableau 4 : Répartition des catégories grammaticales des maillons des chaînes de référence suivant le genre (en %)

[6] Par soucis de lisibilité, ne sont reportés dans ce tableau que les maillons des rangs 1 à 5 (même si les chaînes de référence des *Trois Mousquetaires* ont une longueur moyenne de neuf maillons).

Les disparités observables entre les fréquences des catégories des maillons présentes dans chaque genre textuel sont, dans une certaine mesure, révélatrices des spécificités des chaînes de référence suivant le genre.

- *La classe grammaticale des premiers maillons des chaînes de référence et leur portée :*

Nous nous sommes intéressés ici à définir la catégorie privilégiée des maillons utilisés en première mention (Schnedecker, 1997) dans chacun des genres et à délimiter la portée de ce référent à l'échelle du document. Pour les articles du *Monde*, on relève essentiellement des Np en première mention, avec des redénominations de « Valéry Giscard d'Estaing » dans l'ensemble du document. Ce sont des descriptions définies qui se retrouvent souvent en première mention des chaînes du *Monde Diplomatique*. En effet, parce que les sujets abordés dans les éditoriaux concernent un point de vue à propos des actualités du moment, les auteurs considèrent que les références aux entités présentes dans leurs écrits sont acquises par leurs lecteurs. En cela, la portée de ces référents définis est de l'ordre d'un à deux paragraphes en moyenne.

Les groupes nominaux indéfinis dominent dans les premiers maillons des chaînes de référence de l'*Acquis Communautaire* et des *Trois Mousquetaires*, mais la portée est très locale pour les premières mentions des textes de loi (un paragraphe) alors qu'elle occupe toute une partie du document dans le roman de Dumas (les acteurs sont présentés : « un jeune homme » puis sont mentionnés à de nombreuses reprises dans toute une portion du discours (« le Gascon », « le jeune d'Artagnan », « le don Quichotte de cette autre Rossinante », etc.). Enfin, de même que pour les éditoriaux, La *Documentation Française* compte en majeure partie des GN définis dans les premiers maillons des chaînes (« la satisfaction des clients », « la mesure de la satisfaction des usagers»). Elle en diffère en ce que sa portée concerne l'intégralité du document (portée globale) puisque sont abordés divers aspects à propos du thème central qu'est la satisfaction des clients (la place, la mesure, la recherche de la satisfaction des clients).

- *La correspondance entre la première mention et le thème phrastique :*

Pour ce dernier critère, nous souhaitions savoir dans quelle mesure il était possible de regrouper les chaînes de référence contenant le même thème phrastique. Nous avons donc comptabilisé les cas où le premier maillon des chaînes de référence coïncidait avec le thème de la phrase en cours. On observe ainsi que le premier maillon est le thème phrastique dans 80% des cas pour les articles du *Monde* par exemple, mais qu'il n'est que de l'ordre de 40% pour les rapports issus de la *Documentation Française*.

CORPUS	LONGUEUR MOYENNE
Le Monde	80
Le Monde Diplomatique	100
Acquis Communautaire	60
Les Trois Mousquetaires	60
La Documentation Française	40
moyenne	68

Tableau 5 : Correspondance entre le premier maillon des chaînes et le thème phrastique (en %)

Ainsi, l'étude des chaînes de référence dans un corpus multi-genres a permis de mettre au jour leurs propriétés spécifiques suivant le genre textuel. Ces paramètres seront utilisés pour configurer notre module d'identification des chaînes de référence selon le genre textuel. Par exemple, si le document à traiter est un article de journal, notre système sera paramétré pour identifier des chaînes de référence courtes (d'une longueur moyenne de quatre maillons), qui débuteront de préférence par un Np et dont les maillons seront compris dans la phrase en cours ou dans la phrase suivante. Le Np coïncidera souvent avec le thème phrastique, ce qui signifie que plusieurs chaînes de référence indiqueront le même thème.

Couplés aux autres indices statistiques et linguistiques, les profils-types des chaînes de référence dégagés à l'issue de cette étude en corpus vont permettre à notre système de détecter les thèmes textuels.

4 CONCLUSION ET PERSPECTIVES

Nous avons présenté l'architecture d'un système de détection automatique des thèmes, paramétrable en fonction du genre textuel. Notre méthode hybride statistique / linguistique utilise trois types de marqueurs linguistiques pour attribuer les thèmes aux documents : les marqueurs lexicaux de surface, les marqueurs syntaxiques et les marqueurs référentiels. Nous détaillons les études sur les propriétés de marqueurs référentiels.

L'étude des chaînes de référence dans un corpus multi-genres a permis d'établir une comparaison de ces chaînes suivant cinq critères (longueur des chaînes, distance intermaillonnaire, catégorie grammaticale des premiers maillons, nature des premières mentions, correspondance entre la première mention et le thème phrastique) ; dégageant des spécificités des chaînes de référence suivant le genre textuel. Ces propriétés spécifiques des chaînes de référence vont rendre possible l'adaptation du calcul de la référence suivant le genre textuel du document.

Nous allons enrichir cette étude en corpus en rajoutant d'autres genres textuels (fiches produit) et en étudiant d'autres critères comme les transitions référentielles (la continuité et la discontinuité entre les chaînes de référence). Nous utiliserons les propriétés dégagées dans notre système de calcul des chaînes de référence pour certaines catégories de chaînes de référence dans l'immédiat. Nous projetons enfin de mener une étude sur d'autres catégories de marqueurs (les anaphores nominales), toujours selon le genre textuel, dans le but de détecter des candidats-thèmes.

5 RÉFÉRENCES

Allan J., Carbonell J., Doddington G., Yamron J. et Yang Y. (1998). « Topic Detection and Tracking Pilot Study. Final Report ». Dans *Actes de l'atelier DARPA Broadcast News Transcription and Understanding*.

Ariel M. (1990). *Accessing Noun-Phrase Antecedents*. Londres : Routledge.

Ariel M. (2001). « Accessibility theory: An overview ». Dans T. Sanders, J. Schilperoord et W. Spooren (dir.), *Text Representation*. Amsterdam : Benjamins. p. 29-87.

Biber D. (1994). « Representativeness in corpus design ». *Linguistica Computazionale*, vol. IX-X. Current Issues in Computational Linguistics: in honor of Don Walker, p. 377-408.

Bilhaut F. (2006). *Analyse automatique de structures thématiques discursives - Application à la recherche d'information.* Thèse de Doctorat, Université de Caen.

Charolles M. (1994). « Cohésion, cohérence et pertinence du discours ». *Travaux de Linguistique*, vol. XXIX, p. 125-151.

Charolles M. (1997). « L'encadrement du discours : univers, champs, domaines et espaces ». *Cahier de Recherche Linguistique*, vol. VI, 1-73.

Chastain C. (1975). « Reference and Context ». Dans K. Gunderson (dir.), *Language, Mind and Knowledge.* Minneapolis : University of Minnesota Press.

Choi F. Y. Y., Wiemer-Hastings P. et Moore J. (2001). « Latent semantic analysis for text segmentation ». Dans *Actes de NAACL'01*. p. 109-117.

Corblin F. (1995). *Les formes de reprise dans le discours : Anaphores et chaînes de référence*. Rennes : Presses Universitaires de Rennes.

Cornish F. (1995). « Référence anaphorique, référence déictique, et contexte prédicatif et énonciatif ». Dans Numéro spécial de Sémiotiques, vol. VIII, Anaphores : marqueurs et interprétations. p. 31-55.
Cornish F. (2000). « L'accessibilité cognitive des référents, le Centrage d'attention, et la structuration du discours : une vue d'ensemble ». *Verbum*, vol. XXII, 2000/1.
Daneš F. (1974). « Functional sentence perspective and the organisation of the text ». Dans F. Daneš (dir.), *Paper in Functional Sentence Perspective*. Prague Academia.
Dupont M. (2003). *Une approche cognitive du calcul de la référence*. Thèse de Doctorat, Université de Caen.
Ferret O. et Grau B. (1998). « A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts ». Dans *Actes de ECAI*, Brighton. p. 155-159.
Ferret O. (2006). « Découvrir les thèmes d'un document pour en améliorer la segmentation thématique ». Dans *Actes de la 9ème conférence CIDE*. Fribourg. p. 97-111.
Firbas J. (1992). *Functional Sentence Perspective in Written and Spoken Communication.* Cambridge: Cambridge University Press.
Givón T. (1979). *On understanding grammar.* New York: Academic Press.
Goutsos D. (1997). *Modeling Discourse Topic: sequential relations and strategies in expository text, Advances in Discourse Processes*, vol. LIX. Norwood: Ablex Publishing Corporation.
Grosz B. J., Joshi A. K. et Weinstein S. (1995). « Centering: A framework for modeling the local coherence of discourse ». *Computational Linguistics*, vol. XXI, 1995/2, p. 203-225.
Gundel J., Hedberg N. et Zacharski R. (1993). « Cognitive status and the form of referring expressions », *Language*, vol. LXIX, 274-307.
Gundel J., Hedberg N. et Zacharski R. (2000). « Statut cognitif et anaphoriques indirects », *Verbum*, XXII, 2000/1, 79-102.
Halliday M. A. K. (1985). *An Introduction to Functional Grammar*. Arnold.
Hearst M. A. (1994). « Multi-paragraph segmentation of expository texts ». *Actes de la 32ème rencontre ACL*.
Hernandez N. (2004). *Description et Détection Automatique de Structures de Texte*. Thèse de Doctorat, Université Paris-Sud XI.
Hirst G. et St-Onge D. (1998). « Lexical chains as representations of context for the detection and correction of malapropisms ». Dans C. Fellbaum (dir.), *WordNet: An electronic lexical database and some of its applications*. Cambridge: The MIT Press.
Hobbs J. (1978). « Resolving Pronoun References ». *Lingua*, vol. XLIV, p. 311-338.
Ide N. et Veronis J. (1994). « MULTEXT (Multilingual Tools and Corpora) ». Dans *Actes de la 14ème conference IAACL.* Kyoto.
Illouz G., Habert B., Folch H., Fleury S., Heiden S., Lafon P. et Prévost S. (2000). « TyPTex: Generic features for Text Profiler ». *Content-Based Multimedia Information Access*, vol. II, 1526-1540.
Intuition http://www.sinequa.com/html-fr/fr-edition.oem.html
Ion R. (2007). *TTL: A portable framework for tokenization, tagging and lemmatization of large corpora.* Bucharest: Romanian Academy.
Jenkins C. (2002). « Les procédés référentiels dans les portraits journalistiques ». Dans *Actes du 15ème congrès Skandinaviske romanistkongress*. Oslo.
Kleiber G. (1994). *Anaphores et Pronoms. Louvain-la-Neuve : Duculot.*
Kleiber G. (1999). *Problèmes de Sémantique, Sens et Structures.* Presses Universitaires du Septentrion.
Lambrecht K. (1994). « Information structure and sentence form: Topic, focus, and the mental representation of discourse referents ». *Cambridge Studies in Linguistics*, vol. 71 Cambridge: Cambridge University Press.

Lin C. et Hovy E. (1997). « Identifying topics by position », dans *Actes de la 5 ème conférence ANLP*. p. 283-290.
Lortal G., Todirascu-Courtier A. et Lewkowicz M. (2007). « AnT&CoW:Share, Classify and Elaborate Documents by means of Annotation », *Journal of Digital Information Management*, vol. VI, 2008/1, 61-70.
Marandin J.-L. (1988). « À propos de la notion de thème en discours. Eléments d'analyse dans le récit ». *Langue Française*, vol. LXXVIII, p. 67-87.
Minel J.-L., Descles J.-P., Cartier, E., Crispino G., Ben Hazez S. et Jackiewicz A. (2001). « Résumé automatique par filtrage sémantique d'informations dans des textes, Présentation de la plate-forme FilText ». *Technique et Science Informatiques*, vol. III, Paris : Hermès.
Piérard S. et Bestgen Y. (2006). « Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes ». *TAL*, vol. XLVII, 2006/2.
Popescu-Belis A. (1999). *Modélisation multi-agent des échanges langagiers : application au problème de la référence et son évaluation*, Thèse de Doctorat, Université de Paris XI.
Porhiel S. (2000). « Au sujet de et à propos de – une analyse lexicographique, discursive et linguistique ». *Travaux de linguistique*, vol. XLII/XLIII, p. 171-181.
Porhiel S. (2005). « Les marqueurs de thématisation : des thèmes phrastiques et textuels ». *Travaux de linguistique*, vol. LI, 2005/2, p. 59-88.
Salmon-Alt S. (2001). *Référence et Dialogue finalisé : de la linguistique à un modèle opérationnel*, Thèse de Doctorat, Université H. Poincaré, Nancy.
Schnedecker C. (1997). *Nom propre et chaînes de référence*. Paris : Klincksieck.
Schnedecker C. (2005). « Les chaînes de référence dans les portraits journalistiques : éléments de description », *Travaux de linguistique*, vol. LI, 2005/2, 85-133.
Schnedecker C. (2006). *De l'un à l'autre et réciproquement : aspects sémantiques, discursifs et des pronoms anaphoriques corrélés.* Louvain : De Boeck/université.
Sitbon L. et Bellot P. (2005). « Segmentation thématique par chaînes lexicales pondérées », dans *Actes de TALN 2005*, Dourdan.
Steinberger R., Pouliquen B., Widiger A., Ignat C., Erjavec T., Tufiş D. et Varga D. (2006). « The JRC-Acquis: A multilingual aligned parallel corpus with 20+languages ». Dans *Actes de la 5ème conference LREC*, Italie.
Stokes N., Carthy J. et Smeaton A. F. (2004). « Select: a lexical cohesion based news story segmentation system ». *AI Communications*, vol. XVII, 2004/1, p. 3-12.
Turco G. et Coltier D. (1988). « Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire », *Pratiques*, vol. LVII, 57-79.
Victorri B. (1999). « Traitement automatique des langues et recherche documentaire ». Revue d'interaction Homme-Machine ». prépublication vol. I-II.
Vonk W., Hustinx L. G. et Simons W. H. (1992). « The use of referential expressions in structuring discourse ». *Language and Cognitive Processes*, vol. VII, p. 301-333.

CONFECTION D'UN CORPUS POUR UN NOUVEAU DICTIONNAIRE DE FREQUENCE DU FRANÇAIS

Deryle Lonsdale et Yvon Le Bras
Brigham Young University

À l'heure actuelle, le français en tant que langue étrangère est la seconde langue enseignée dans le monde juste après l'anglais. Chose étonnante, il n'existait jusqu'à la publication de notre dictionnaire en avril 2009 aucun autre dictionnaire de fréquence basé sur un large corpus qui permette d'en faciliter l'apprentissage.

Les seuls dictionnaires de fréquence du français dignes de ce nom sont ceux d'Henmon (1924) et de Juilland (1970), basés respectivement sur des corpus de 400 000 et 500 000 mots. L'information relative aux mots qu'ils contiennent demeure minimale cependant. Si la liste des 3 500 mots du français fondamental de Gougenheim (1958) s'accompagne bien de définitions, le fait qu'elles sont formulées exclusivement en français en rend la compréhension difficile pour les apprenants non-francophones. Quant aux listes de vocabulaire de Brunet (1981) développées à partir des données du *Trésor de la Langue Française* (Imbs, 1971-1994) ou de Beauchemin (1992), tirées d'un corpus du français parlé au Québec, elles s'adressent essentiellement à un public de spécialistes.

À l'exception de quelques dictionnaires pour débutants publiés par Oxford University Press (Corréand, 2006), Living Language (Lazare, 1992) et Dover Publications (Buxbaum, 2001) qui proposent une liste des 1001 à 20 000 mots français « les plus utiles », sans toutefois nous éclairer sur la façon selon laquelle ils ont été sélectionnés, les seules ressources lexicales disponibles sont celles de ARTFL FRANTEXT et TLFI sur l'internet, bien que les frais d'abonnement et l'accès en ligne n'en favorisent guère l'usage.

Il importe de noter ici que notre dictionnaire est le dernier né d'une série de dictionnaires de fréquence publiés ces dernières années par Routledge (Taylor & Francis Group) et dont les auteurs sont, à l'exception du dictionnaire chinois, de Brigham Young University, Provo, Utah, USA. Vous pouvez d'ailleurs consulter les corpus qui ont servi à confectionner la plupart de ces dictionnaires sur le site corpora.byu.edu.

1 LE CORPUS

Notre dictionnaire se base sur un corpus d'environ 23 millions de mots, la moitié étant constituée de données orales et l'autre de textes écrits. Nous avons choisi ce chiffre parce qu'il correspond à peu près à la taille de l'« American National Corpus », célèbre corpus de l'anglais américain qui compte 22 millions de mots.

Pour constituer ce corpus nous nous sommes reposés dans une certaine mesure sur ce qui a déjà été fait dans ce domaine, en nous servant de ressources du domaine public. Nous avons donc sélectionné et téléchargé un certain nombre de documents de sites

Internet établis par des organismes gouvernementaux, des sociétés et associations littéraires. Nous nous sommes également procurés des corpus oraux et écrits déjà disponibles et en avons extrait les données nécessaires. Ce faisant, nous avons veillé à respecter tous les droits de reproduction et d'utilisation appropriées des dites ressources.

Nous n'avons pas cherché à proportionner de façon exacte son contenu d'après une représentation démographique du monde francophone. Nous n'y avons inclus, par exemple, que très peu de textes africains, tahitiens, antillais, etc. Toutefois, on y trouve une bonne représentation de données originaires du Canada ainsi que de tous les pays francophones d'Europe.

Comme nous n'avons pas non plus voulu faire une analyse diachronique du vocabulaire français, aucun des documents dont nous sommes servis n'ait antérieur à 1950. Il existe un équilibre approximatif entre les genres de textes recueillis. Notre représentation des genres oraux et écrits n'est cependant pas aussi variée que dans d'autres corpus, comme par exemple le British National Corpus. Nous avons recueilli des textes dans leur intégralité et des fragments ou échantillons des textes modernes choisis au hasard.

La partie orale du corpus comprend des documents – extraits ou textes intégraux – de plusieurs sortes : procès-verbaux de réunions, conférences et débats ; transcriptions d'appels téléphoniques et de discussions ; transcriptions d'interviews avec des personnalités en vue tels comme des écrivains, des hommes et femmes d'affaires, des athlètes, des universitaires, des représentants des médias et du monde du spectacle. Nous avons aussi inclus à ce corpus oral des textes de fiction comme des scénarios de film et des sous-titres, ainsi que des dialogues de théâtre.

La moitié du corpus écrit est composée d'articles d'agences de presse, de quotidiens, d'hebdomadaires et de mensuels ; de magazines de vulgarisation scientifique et technique; de même que des communiqués de presse, de la correspondance commerciale et des chapitres de manuels techniques. Il comprend aussi une section littéraire où l'on trouve des textes de fiction comme des romans et des nouvelles, ainsi que des mémoires et des essais.

La composition du corpus se présente comme suit :

	# de mots (approx.)	Genre
Oral		
	175,000	Conversations
	3,750,000	Hansard canadien
	3,020,000	Transcrits d'interviews
	1,000,000	Débats parlementaires de l'UE
	855,000	Appels téléphoniques
	470,000	Dialogues de théâtre
	2,230,000	Sous-titres de films
TOTAL	11,500,000	
Écrit		
	3,000,000	Agences de presse
	2,015,000	Articles de quotidiens
	4,734,000	Œuvres littéraires(fiction, non-fiction)
	434,000	Magazines de vulgarisation technique
	1,317,000	Bulletins de presse, manuels d'emploi
TOTAL	11,500,000	
AU TOTAL	23,000,000	

Tableau 1 : Composition du corpus

Nous exposerons maintenant en détail le traitement du corpus et l'identification du vocabulaire qui nous intéressait, à savoir les 5000 mots les plus fréquents du français moderne.

2 LE TRAITEMENT DU CORPUS

Une fois les données du corpus rassemblés, nous avons entrepris d'en faire l'annotation. Il nous a d'abord fallu standardiser la structure et l'encodage du contenu des documents, qui se présentaient dans une grande variété de formats et de caractères (EBCDIC, MACROMAN, ISO, UTF-8 et HTML, entre autres). Nous avons dû également effacer toutes données extra-linguistiques : images, annonces publicitaires et codes de formatage.

Ce travail s'est fait par le biais de plusieurs techniques de traitement des langues naturelles et de plusieurs outils de programmation tels que des scripts Perl, des ressources Unix (make, awk, grep, sort, uniq, join, comm) et des parseurs SGML, HTML et XML.

Ceci nous a permis de faire la tokenisation de tous les documents, c'est-à-dire l'identification des mots et de la ponctuation. Cette tâche s'est avérée plutôt difficile étant donné l'ambiguïté fonctionnelle des apostrophes et des traits d'union (l'homme vs. aujourd'hui, dis-moi vs. week-end). Dans certains documents, les lettres majuscules portaient des accents qui manquaient dans d'autres. Il nous a fallu aussi standardiser plusieurs symboles représentant la ponctuation (les points de suspension, le tiret, les guillemets), la typographie (le degré, les puces) ainsi que les symboles mathématiques et monétaires.

Nous avons procédé ensuite à l'analyse morphologique du corpus en assignant à chaque mot une étiquette qui correspond à sa catégorie lexicale (nom, verbe, adjectif) et nous avons en calculé le (ou les) lemme(s) possible(s). Cette analyse s'est faite d'une

manière automatique à l'aide d'un certain nombre de logiciels à la disposition des chercheurs. Pour ce faire, nous avons élaboré notre propre jeu d'étiquetage.

Ensuite nous avons effectué la lemmatisation de notre corpus, c'est-à-dire à ne retenir que la forme canonique des lemmes, ou mots (pour un verbe : ce verbe à l'infinitif et pour les autres mots, le mot au masculin singulier). Il va sans dire que la lemmatisation, essentielle pour calculer la fréquence des mots, est une tâche ardue (formes non-définies, ambiguïté morphologique, abréviations, majuscules/minuscules, symboles) qui a été facilitée grâce à l'utilisation de bases de données lexicales très utiles pour le français comme FrnWordNet et BDLEX.

3 L'ANALYSE STATISTIQUE DU VOCABULAIRE

Le traitement linguistique du corpus terminé, nous sommes passés à l'indentification du vocabulaire cible, c'est-à-dire aux 5 000 mots qui y apparaissent le plus souvent. Nous avons combiné toutes les différentes variations d'un mot pour calculer le nombre de ses occurrences. Par exemple, le montant total de toutes les occurrences du mot « pour », préposition ou conjonction, en lettres majuscules ou minuscules, est 151 709. Pour les noms, verbes et adjectifs il a fallu tenir compte bien sûr des variations dérivationnelles (malheur/malheureux) et flexionnelles (conjugaison, pronoms personnels, singulier/pluriel, masculin/singulier) des lemmes. Dans notre corpus, on trouve 25 différentes formes du verbe « déterminer » ; toutes ces formes contribuent au montant total des occurrences de ce mot. Nous avons néanmoins supprimé les symboles et les noms propres du corpus avant de faire ces calculs.

Étant donné la présence de différents genres et quantités de textes dans notre corpus, nous avons jugé qu'un simple calcul de fréquence ne suffirait pas à identifier les mots les plus usités. Il s'agit là d'un problème auquel sont souvent confrontés les linguistes de corpus : si un mot s'avère très fréquent dans une « tranche » du corpus mais assez rare dans une autre, un calcul basé uniquement sur la fréquence risque d'être peu satisfaisant. Des dizaines de formules statistiques sont utilisées pour remédier à ce problème, mais elles présentent toutes des inconvénients. Une des formules les plus récentes et les plus prometteuses s'appelle la « déviation des proportions » ('deviation of proportions'), ou DP (Gries, 2008), et c'est cette métrique que nous avons choisie pour circonscrire notre vocabulaire.

La métrique DP vise à mettre en évidence la proportion des occurrences d'un mot à travers les différentes « tranches » du corpus. Le calcul de chaque mot se base sur : (i) la « proportion observée », c'est-à-dire le montant total de toutes les occurrences de toutes les formes du mot dans une tranche, ensuite normalisant ce chiffre par la fréquence du mot sur le corpus entier ; (ii) la « proportion attendue », c'est-à-dire la normalisation de chaque « tranche » du corpus par rapport au grandeur total du corpus ; et (iii) le calcul du mesure DP, qui est la somme de toutes les différences absolues entre les proportions observées et attendues, divisé par deux. Il en résulte une métrique entre 0 et 1 : 0 représentant une distribution parfaitement uniforme du mot à travers le corpus tout entier et 1, un mot dont la dispersion est restreinte à une tranche du corpus.

Bien qu'elle soit très utile pour montrer la distribution des mots à travers le corpus, il nous a fallu pour établir un ordre cohérent des mots dans notre dictionnaire tenir compte à la fois de la « déviation des proportions » et de la fréquence brute des mots. C'est ce que nous avons fait en divisant la fréquence de chaque lemme par l'indice de déviation

des proportions correspondant. Nous avons ensuite fait le calcul mathématique de l'indice de dispersion des mots à l'aide de la formule suivante ($100*\exp^{-DP)}$. Les valeurs qui s'approchent de 100 indiquent que le mot est distribué de façon égale à travers tout le corpus ; les valeurs en-dessous de 50 indiquent que les mots ne se trouvent que dans des parties restreintes du corpus.

Bien que ces calculs soient quelque peu complexes, il importe de garder à l'esprit que les mots dans ce dictionnaire sont classés en fonction de la fréquence totale de leurs différentes formes et de leur dispersion/distribution à travers les différentes parties du corpus.

4 LES ENTRÉES

Il s'agit là de la partie principale du dictionnaire. On y trouve la liste des 5000 lemmes français les plus fréquents par ordre numérique, en commençant par le plus fréquent. Chaque entrée comprend les données suivantes : ordre numérique (1, 2, 3…), lemme, partie(s) du discours, terme(s) équivalent(s) en anglais, contexte d'usage tiré du corpus, traduction en anglais du contexte, chiffre de dispersion, fréquence brute, code de variation de registre. Par exemple, l'entrée pour le mot « aimer » se présente ainsi:

242 aimer v to like, love
tu sais que je t'aime
you know I love you
71 | 10085 -n

Cette entrée montre que le lemme en question (et ses variantes morphologiques) est classé 242ème parmi tous les mots français en fonction de sa fréquence brute et de sa dispersion dans le corpus. L'étiquette indique que le mot est un verbe. Deux traductions possibles en anglais américain sont « to like » et « to love ». Ces traductions ne sont que représentatives, toutes les possibilités de traduction n'étant pas mentionnées. Un contexte tiré du corpus où l'on trouve une des formes conjuguées du verbe « aimer » apparaît ensuite, accompagné de sa traduction en anglais. Le chiffre « 71 » indique la valeur de dispersion du mot dans le corpus sur une échelle de 27 à 100 ; le mot se présente donc d'une manière régulière dans tout le corpus. Le chiffre « 10085 » indique également sa fréquence brute (nombre d'occurrence dans le corpus). Finalement, un code de registre **–n** signale que ce mot est moins fréquent que prévu dans les textes non littéraires (non fiction).

5 LE TRAITEMENT DES CONTEXTES

Nous avons mis au point un logiciel particulier afin d'extraire du corpus les contextes de toutes les formes des mots du dictionnaire. Nous avons ensuite réduit leur nombre (une douzaine environ par entrée) à l'aide d'un autre programme informatique en fonction de leur brièveté, de leur autonomie syntaxique et sémantique (proposition ou phrase complète) et de leur contenu lexical correspondant au vocabulaire cible.

Le choix du meilleur contexte a été un travail long et difficile étant donné le grand nombre de contextes à considérer. Il a fallu faire preuve d'intuition ici, trouver des contextes brefs et concis, au sens évident, tout en réfléchissant à leur traduction possible

en anglais. Dans la liste suivante, on voit des exemples de contextes utiles (✓) et peu utiles (✕) pour les mots soulignés :

✓ les Bulls, qui comptent désormais quatre victoires
✕ il se le fût désormais tenu pour dit

✓ ils m'ont posé des questions. des tas de questions.
✕ la question se pose de manière inverse!

✓ la mise en place progressive d'un nouveau système
✕ aucune mise aux enchères prochaine n'est prévue

✓ j'ai soulevé des questions assez simples
✕ dieux anciens, rois médiévaux ou simples esclaves, tous sont bien vivants

Chaque contexte sélectionné a été traduit comme il se doit en anglais. Quoiqu'un certain nombre de ces contextes proviennent de documents déjà traduits, nous n'avons pas utilisé ces traductions pour éviter d'introduire des concepts extérieurs tirés de contextes plus larges. Nous avons donc considéré chaque contexte individuellement et l'avons traduit essentiellement à la main en faisant parfois appel à des ressources en ligne.

Comme il s'agissait de se focaliser davantage sur le sens du contexte que sur sa structure, son style ou son registre, sa traduction en anglais ne lui correspond pas toujours exactement du point de vue lexical. Par exemple, le terme anglais apparaissant devant chaque entrée du dictionnaire n'est pas nécessairement celui qui a été retenu dans la traduction du contexte, bien que leur sens soit équivalent. D'autre part, comme chaque contexte est détaché du document d'où il provient, le genre des personnes et des choses auxquelles les pronoms et articles se rapportent n'est pas toujours évident. Ainsi, un contexte comme « je lui téléphone » pourrait se traduire soit « I'm telephoning him » soit « I'm telephoning her ». Notre intention étant de fournir aux utilisateurs non francophones de notre dictionnaire une traduction aussi compréhensible que possible du contexte en anglais, cela ne constitue aucun inconvénient à nos yeux.

Afin de rendre l'usage du dictionnaire plus aisé pour le lecteur, nous avons créé des tableaux thématiques de vocabulaire utile (p. ex. : le corps, la nourriture, le temps, verbes de mouvement, etc.). D'autres tableaux donnent par exemple des renseignements sur l'utilisation du pronom « se », la fréquence par rapport à la longueur des mots, les variations d'usage des mots en fonction du registre de langue, etc. Voir à ce propos le tableau suivant :

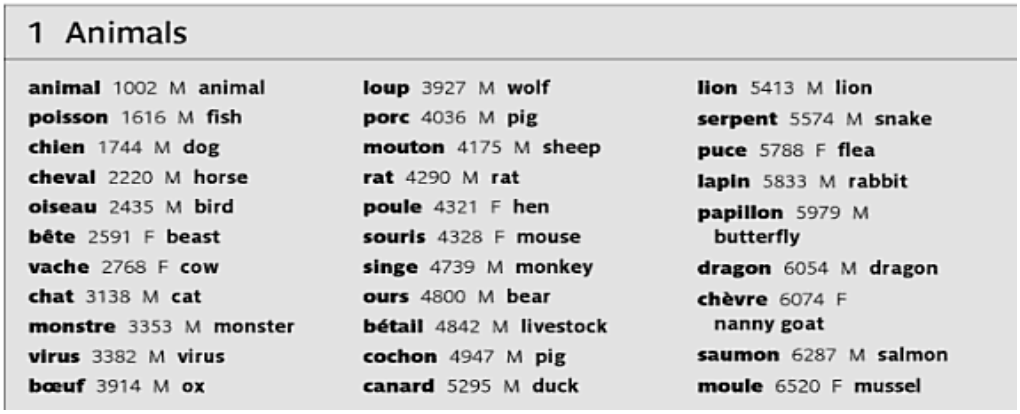


Figure 1 : Liste thématique

6 CONCLUSION

Comme il s'agissait ici de créer un dictionnaire de fréquence lexicale proprement dit, on n'y trouve aucune transcription phonétique des mots, ni de référence à leur étymologie ou à leur domaine d'usage. Dans le même ordre d'idée, nous n'avons pas cherché à y répertorier les locutions figées, collocations ou expressions idiomatiques présentes dans le corpus, ce qui pourrait d'ailleurs faire l'objet d'un projet ultérieur. Ainsi, ce dictionnaire n'est ni plus ni moins qu'un outil de référence qui, nous l'espérons, permettra aux étudiants d'acquérir dans les meilleures conditions un vocabulaire de base du français moderne.

7 RÉFÉRENCES

Beauchemin N., Margel P. et Théoret M. (1992). *Dictionnaire de fréquence des mots du français parlé au Québec: fréquence, dispersion, usage, écart réduit.* New York : P. Lang.

Brunet É. (1981). *Le vocabulaire français de 1789 à nos jours d'après les données du Trésor de la langue française*. Paris : Champion.

Buxbaum M. O. (2001). *1001 Most Useful French Words*. Mineola, NY : Dover Publications.

Corréand M. (2006). *New Oxford Beginner's French Dictionary.* New York, NY : Oxford University Press.

Davies M. (2006). *Frequency Dictionary of Spanish: Core Vocabulary for Learners*. New York, NY : Routledge.

Davies M. et Preto-Bay A. M. R. (2008). *Frequency Dictionary of Portuguese: Core Vocabulary for Learners*. New York, NY : Routledge.

Gougenheim G. (1958). *Dictionnaire fondamental de la langue française.* Paris : Librairie Marcel Didier.

Gries S. T. (2008). « Dispersions and adjusted frequencies in corpora ». *International Journal of Corpus Linguistics*, vol. 13, no. 4, p. 403-437.

Henmon V.A.C. (1924). *A French word book based on a count of 400,000 running words.* Madison, WI : University of Wisconsin.

Imbs P. (1971-1994). *Trésor de la langue française.* Paris : CNRS, Gallimard.
Juilland A., Brodin D. et Davidovitch C. (1970). *Frequency Dictionary of French Words*. Mouton.
Lazare L. (1992). *French Learner's Dictionary.* New York, NY : Living Language.

TRAITEMENT LEXICOGRAPHIQUE DE L'EMPRUNT DANS UN CORPUS DE DICTIONNAIRES BILINGUES DE LA PÉRIODE COLONIALE FRANÇAISE EN ALGÉRIE

Mahfoud Mahtout
Laboratoire LiDiFra – Université de Rouen

1 INTRODUCTION

Le traitement lexicographique de l'emprunt pendant la période coloniale requiert une attention toute particulière pour des raisons multiples. D'abord, il existe une quantité non négligeable de mots français empruntés aux langues locales d'Algérie- essentiellement à l'arabe dialectal et au kabyle- et inversement. En outre, les langues d'Algérie n'appartiennent pas à la famille indo-européenne mais à la branche sémitique des langues chamito-sémitiques. De plus, la vague d'emprunts propres à la période coloniale est passée sans intermédiaire des langues locales au français et vice-versa. En effet, cette période coloniale est marquée par la coexistence des communautés européenne et indigène sur le même sol. Des rapports de communication quotidienne favorisent un échange bilatéral entre les deux communautés et font naître de nouveaux besoins pour désigner la nouvelle réalité coloniale. Même si cette situation est à caractère conflictuel, l'emprunt trouve sa place pour assurer l'intercompréhension entre les différentes communautés linguistiques et couvrir les notions qui n'ont pas d'équivalents dans l'une ou l'autre des langues en contact. Quant aux langues locales, elles ne sont pas équipées pour exprimer les nouvelles notions introduites par la colonisation. Elles recourent à l'emprunt pour répondre à la nécessité de pouvoir combler un vide, exprimer des notions jusque là inconnues des locuteurs des langues locales.

1.1 Problématique

Il ne saurait être question ici de faire l'histoire linguistique des emprunts entre le français, l'arabe et le kabyle ni de démontrer leur importance dans l'ensemble des dictionnaires de la période coloniale. Nous nous proposons plutôt de rendre compte du phénomène de l'emprunt du français vers l'arabe dialectal algérien et le kabyle et inversement. Le passage d'une unité linguistique d'une langue source à une langue cible suppose quelques modifications nécessaires à son intégration dans le système linguistique de la langue emprunteuse. C'est ce processus progressif d'intégration des emprunts et son traitement lexicographique que nous proposons d'aborder dans cette étude. Nous nous posons les questions suivantes : Quel processus d'intégration et d'adoption les emprunts suivent-ils en français, en arabe dialectal et en kabyle ? Quel est leur traitement lexicographique dans les dictionnaires bilingues et dans le TLFi ? L'arabe dialectal et le kabyle traitent-il de la même manière les unités empruntées communément au français ? Nous allons essayer de répondre à ces questions en nous basant sur des critères linguistiques.

Pour nous rendre compte du traitement lexicographique des emprunts coloniaux, nous avons choisi deux dictionnaires bilingues de la période coloniale : le *Dictionnaire français-*

kabyle[22] et le *Dictionnaire français-arabe de la langue parlée en Algérie*[23]. Ces dictionnaires recueillent des faits établis et confirmés par l'usage, nous offrent un corpus riche et varié en données lexicographiques. Par ailleurs, pour comprendre avec un recul historique les stratégies d'admission des unités empruntées, nous avons consulté le *Trésor de la langue française informatisé*, désormais *TLFi*. Il nous permet, en effet, de retracer l'histoire des mots étrangers du XIX^{e} et XX^{e} siècle grâce à ses notices historiques et étymologiques qui nous offrent des indications concernant la prononciation, la graphie retenue, le genre, le nombre, etc, des lexies empruntées.

1.2 Quelques domaines d'emprunt

Les emprunts retenus sont ceux que l'arabe dialectal et le kabyle ont faits au français et inversement, recueillis dans les deux dictionnaires bilingues mentionnés ci-dessus. Les emprunts de la période coloniale couvrent plusieurs domaines. Le français a emprunté à l'arabe dialectal et au kabyle des termes qui dénotent une réalité propre à l'environnement social et culturel qui caractérise la société traditionnelle algérienne tels que : *goum*, *cadi*, *gandoura*, *turban*. D'autres font référence au domaine religieux tels que : *cheikh, iman*, *mufti*, *zaouïa*. D'autres, enfin, renvoient à l'univers militaire et à l'organisation administrative de la colonie tels que : *razia*, *caïd*, *agha*, *amin*.

Quant à l'arabe dialectal et le kabyle, ils ont empruntés plutôt des termes pour qualifier des objets tels que : fourchette, *ferchita*, allumette, *zalamit*, casserole, *takasrunt* etc, et beaucoup d'autres termes pour désigner la nouvelle organisation militaire et administrative de la colonie. Nous retrouvons des mots tels que : officier, *fecian*, sergent, *sarjan*, consul, *qounçou,* école, *lakul* etc.

Nous constatons à travers ces quelques exemples que les langues en contact se sont adaptées à cette nouvelle situation d'intercommunication en empruntant mutuellement des termes pour désigner des notions inconnues ou peu connues, des nouvelles réalités et renforcer ainsi le champ référentiel des locuteurs.

1.3 Traitement linguistique des emprunts coloniaux et leurs procédés d'intégration dans les dictionnaires.

Il est clair que l'adoption de l'emprunt dans la langue d'accueil est conditionnée par l'usage et que les langues emprunteuses habillent les mots en fonction des contraintes imposées par leurs systèmes : phonologique, morphologique, syntaxique et sémantique. Les changements subis par l'emprunt sont à observer du point de vue du français, de l'arabe et du kabyle.

Quant aux dictionnaires bilingues de la période coloniale, ils réservent un traitement lacunaire aux emprunts. Hormis les très rares mentions sur l'origine du terme, leurs auteurs se contentent de mentionner le mot emprunté et parfois son pluriel, sans aucune autre indication permettant à l'usager de mieux comprendre son emploi, son sens et sa dérivation, le cas échéant. L'utilisateur est donc contraint de consulter un autre dictionnaire pour déterminer l'origine du terme, sa forme initiale, le sens que le mot avait avant d'être admis dans le vocabulaire de la langue emprunteuse.

Prenons des termes tels que *karrosse*, *cart'a*, *loutil*, *takasrut*, *acanti* pour (carrosse, carte, hôtel, casserole, chantier), ces emprunts admis dans les dictionnaires de la période coloniale

[22] HUYGHE, G, Père Blancs, *Dictionnaire français-kabyle, Qamus Rumi-Qbaili*, France, L. & A. Godenne, 1902-1903, 22cm, 893 p.

[23] BEN SEDIRA Belkassem, *Dictionnaire français-arabe de la langue parlée en Algérie* , Cinquième édition, Alger, Adolphe Jourdan, 1886, In-16, 790 p.

montrent qu'ils sont largement utilisés par les locuteurs des langues locales. En outre, ces termes du XIX^{e} et du début du XX^{e} siècle ont réussi à traverser les époques pour être, encore aujourd'hui, utilisés par les locuteurs algériens. Cette traversée des siècles se mesure par l'intégration profonde des termes dans le nouveau système linguistique de la langue d'accueil.

2 AU NIVEAU PHONÉTIQUE ET PHONOLOGIQUE

L'admission d'un terme français dans la langue arabe ou kabyle subit les règles phonétiques et phonologiques de ces langues d'accueil. En outre, *la structure syllabique* du français diffère de celle de l'arabe et du berbère. En arabe dialectal ainsi qu'en kabyle, les mots sont constitués, d'une manière générale, de deux, parfois, de trois syllabes. À ce propos, Zakia Iraqui Sinaceur (2004 : 510), nous fait remarquer que les mots français qui contiennent plus de deux syllabes ont subi différents traitements phonologiques qui les ont ramenés au modèle du substantif arabe. Dans les exemples suivants nous signalons les procédés les plus fréquents. Le traitement d'unités comme : ennuyer ; *nouitini*, officier ; *fecian,* résulte d'une aphérèse qui consiste à supprimer la syllabe initiale. Les emprunts : numéro ; *noumro* ; capitaine, *quobt'an*, sont traités par syncope, procédé qui consiste en la suppression d'une syllabe à l'intérieur du mot. Les emprunts comme : as ; *las*, hôtel ; *loutil*, école ; *lakul*, suivent un processus d'agglutination qui permet, dans ce cas, de réunir un article et un substantif. En outre, les mots comme : carte ; *cart'a*, bière ; *birra,* subissent un allongement par rajout d'une voyelle finale.

2.1 La structure vocalique

Les emprunts subissent des modifications dictées par la structure vocalique de la langue emprunteuse. Il se trouve que l'arabe dialectal et le kabyle ne disposent que de trois voyelles fondamentales [(a, i et u), plus « e » pour le kabyle appelé *schwa* ou voyelle zéro, et « ə », semi-voyelle, pour l'arabe dialectal], pour intégrer les termes français plus riches au niveau vocalique. De ce fait, les locuteurs de l'arabe dialectal et du kabyle adaptent les unités empruntées au vocalisme déjà connu dans leurs langues. Observons les exemples suivants :

- Le phonème [**o**] en français se réalise en [**u**] dans : carrosse : *karrouça* ; bateau : *babur* ; casserole ; *takasrut* ou *takasrunt*.
- Les phonèmes [**e**] et le [**ɛ**] se transforment [**i**] dans : timbré : *tenbri* ; géomètre : *jiomitr* ; fourchette ; *ferchita* ; hôtel : *loutil*.
- Le phonème [**y**] se transforme en [**u**] dans : numéro : *noumro* ; écurie : *kuri*

Nous constatons à travers ces quelques exemples que malgré les contraintes imposées par le système vocalique de l'arabe dialectal et du kabyle, les emprunts français se sont adaptés à celui-ci en subissant des transformations phonologiques. Ces transformations sont autant d'indices nous renseignant sur le degré d'intégration des unités empruntées.

La dénasalisation

Comme beaucoup d'autres langues, l'arabe dialectal et le kabyle ne possèdent pas de voyelles nasales. Les locuteurs ont donc tendance à dénasaliser les voyelles françaises et à les transformer en voyelles orales simples. En outre, nous remarquons que cette réduction de voyelles nasales s'accompagne souvent par une adjonction de [**n**] en appendice.

- Les nasales [**ɑ̃**] se transforme en **/a+n/** dans : sergent : *sarjan* ; chantier : *acanti* : gendarme ; *ajandarmi*
- [**ɔ̃**] se transforme en **/u+n/** dans : bataillon : *abataiun* ; bidon : *abidun* ; savon : çaboun

- [æ̃] se transforme soit en /a+n/ ou /e+n/ dans : américain : mrikan ; timbré : tenbri ; magazin : makhzen

Le fait que les usagers de l'arabe dialectal et du kabyle opèrent une dénasalisation des voyelles nasales françaises peut signifier que l'emprunt est bien intégré dans la langue d'accueil. Par ailleurs, cette opération de dénasalisation permet d'opérer une harmonie vocalique entre les mots empruntés et ceux de la langue source, en leur faisant appliquer les mêmes principes de réalisation phonétique.

2.2 La structure consonantique

Si l'arabe et le kabyle sont des langues à vocalisme pauvre, leurs systèmes consonantiques sont très riches. Outre les consonnes emphatiques, l'arabe dialectal et le kabyle possèdent des phonèmes à articulation très en arrière tels que les vélaires, les laryngales et pharyngales. Cette richesse offre aux emprunts français une grande souplesse d'intégration dans la langue d'accueil. Néanmoins, si nous comparons le système consonantique français à celui de l'arabe et du kabyle, nous nous apercevons que les consonnes [p] et [v] ne figurent pas dans ceux-ci et présentent donc une case vide. Les emprunts vont tout de même s'adapter à la structure consonantique de l'arabe et du kabyle et les consonnes manquantes vont être remplacées par celles qui se trouvent au même point d'articulation.

28. L'occlusive bilabiale sourde [p] se réalise par la sonore [b] dans : capitaine : quobt'an ; espagnol : sbaniul ; lampe ; lamba
29. La fricative labiodentale sonore [v] se réalise généralement par [b] dans : savon ; çaboun : taverne ; teberna, mais aussi par [f] dans zouave ; zuaf

Voyons maintenant comment s'adaptent les emprunts français, faits à l'arabe dialectal et au kabyle, sur le plan consonantique.

Nous avons vu que le système consonantique de l'arabe dialectal et du kabyle est plus riche que celui du français. De ce fait, les emprunts subissent des modifications qui leur permettent de s'adapter à la structure consonantique du français. Observons les quelques exemples suivants recueillis dans les deux dictionnaires bilingues cités ci-dessus :

2. L'uvulaire occlusive sourde *qâf* **[ق]**[24] se transforme souvent en [**k**] dans : cadi : **قاضي** ; café : **قهوة** ; caïd : **قايد** ; couffin : **قفة** ; cafetan : aqef'dan. Cependant, le **[ق]** peut se réaliser par la post-palatale occlusive sonore *gâf* [**ڤ**] qui correspond à la vélaire occlusive sonore française [**g**] dans : gandoura : **قندورة** [gɑ̃duRa] ; goum : **قوم** [gum]

Observons maintenant quel traitement lexicographique le TLF*i* réserve à l'emprunt *gandoura* dans sa notice historique et étymologique:

Le TLF*i* commence d'abord par donner des indications sur l'orthographe du substantif *gandoura(h)* qui s'écrit avec deux graphies différentes : *Gandoura* ou *Gandourah*. Ensuite, il nous fournit un élément de datation historique du terme *gandoura* qui remonte à 1756 correspondant à l'époque de la concession française en Algérie (la Calle). Le TLFi nous offre des informations étymologiques du mot *gandoura* emprunté à l'arabe dialectal algérien, mais déjà présent dans le moyen français sous forme de *arcandore*. Enfin, la transcription phonétique du mot gandoura est fourni en arabe dialectal algérien [gandūra] et en arabe littéral [qandūra] pour aider, sans doute, l'usager à différencier les deux réalisations du terme selon les besoins dictées par la situation de communication (formelle *vs* informelle).

3. *L'uvulaire fricative sourde* [**خ**] représentée par [χ] se transforme le plus souvent en [**k**] dans : calife : **خليفة** [kalif], pl. *khoulfaouat*, **خلفوات**, *khoulafa*, **خلفاء** ; cheikh : **شيخ** [ʃɛk] pl. achiakh, **اشياخ** ou mechaïkh, **مشايخ** . Nous retrouvons les mêmes emprunts dans le

[24] Nous avons gardé les même lettres alphabétique telles quelles sont présentées dans le *Dictionnaire français-arabe de la langue parlée en Algérie*.

dictionnaire français-kabyle mais avec une translittération différente: calife : *ħalifa* ; cheikh : *ciħ*, pl. *mcaiħ*. Toutefois, le [**خ**] se réalise aussi par la vélaire occlusive sonore [**g**] dans : magazin : **مخزن** [maḥzan] *makhzen*, pl. **مخازن** ; *mekhazen*. Par ailleurs, il y a lieu de remarquer que le phonème [**خ**] est rendu le plus souvent en français par deux graphies différentes : ***c*** comme dans **c**alife, ***kh*** comme dans chei**kh**.

Le TLF*i* a admis l'emprunt *cheik* sous deux graphies différentes : *cheikh* et *cheik*. Le dictionnaire signale également une variation au niveau de la prononciation du mot *cheikh* qui varie entre [ʃɛk] et [ʃɛjk] accompagné d'attestations d'usage dans certains ouvrages de référence. La notice historique nous offre une première datation du terme qui remonte à 1309 sous cette forme « *seic* » au sens de *chef de tribu chez les arabes* puis évolue pour devenir, en 1631, *cheik* au sens de *prêtre de la religion musulmane*. L'étymologie du mot *cheikh* met en évidence le mot venant de l'arabe [šayḫ] qui désignait un *vieillard*. Dans le dictionnaire français-arabe de Ben Sedira, le mot *cheikh* est "contextualisé" par (*d'une ville*) **شيخ البلد** translittéré *cheikh el blad*, et par (*d'une tribu*) **شيخ العرش** translittéré *cheikh el árch*. C'est deux contextes nous laissent penser à deux significations différentes du substantif *cheikh*. Dans l'idée où la *ville* est un centre urbain et une agglomération relativement importante en terme d'habitants et d'habitations, le mot *cheikh* peut signifier *imam* "chef de prière dans une mosquée". Cependant, si le sens de la tribu (cheikh *d'une tribu*) se rapproche de celui d'un groupe social fondé sur une parenté ethnique, unis par des règles et des traditions, le mot *cheik* peut signifier le titre donné à un homme respectable par son âge, ses connaissances et son expérience.

4. *L'uvulaire fricative sonore* [**غ**] se réalise soit par [**R**] soit par [**g**] dans : razzia[25] : **غازية**, *r'azia*, pl. **غوازي** : *r'ouazi* ; gazelle : **غزالة**, *r'ezala*, pl. **غزلان**, *r'ezlan*, collecitif, **غزال**, *r'ezal*.

Le TLF*i*, dans son traitement lexicographique du mot *gazelle*, commence par donner une précision sur la prononciation du terme qui variait autrefois entre [ġazɛl] et [g□zel] et souligne que la seconde est de plus en plus rare. Quant à la notice historique, elle atteste de la présence du terme dans la langue française depuis 1195 sous la forme de *gacele*, et de sa présence en latin médiéval sous forme de *gazela*. Le TLF*i*, note que l'origine étymologique du mot *gazelle* vient de l'arabe classique [ġazāl] et de l'arabe maghrébin [ġazēl]. Notons que la prononciation de cette dernière est la plus proche de la réalisation française du mot *gazelle*. Cependant, le TLF*i* ne fournit aucune indication ni attestation sur l'emprunt du mot *gazelle* à l'arabe dialectal maghrébin.

5. *L'uvulaire fricative sonore* [**ع**] rendu par en français par [**a**] dans : salamalec : **سلام عليك**, *salam álik* ; ambre, **عنبر**, ánber ; arabe, **عربي**, a'rbi, pl. **عرب**, a'rab. Le TLF*i* fait remonter l'origine de ces termes à l'arabe et au monde musulman sans préciser de quel arabe il s'agissait. Cependant, les termes évoqués font partie de la nomenclature du dictionnaire français-arabe de la langue parlée en Algérie de Ben Sedira.
6. *L'interdentale fricative sonore* [**ذ**] se rend parfois par [**z**] dans : muezzin : *muddin*. Le TLF*i*, présente deux prononciations différentes de ce terme [mʮɛdzin] et [mʮᵉ(d)zɛ̃].
7. *La semi-voyelle labialo-vélaire* [**و**] se réalise parfois par [**w**] et d'autre fois par [**v**] dans : zaouïa : **زاوية**, *zaouïa* ; vizir : **وزير** *ouzir* ; zouave ; *zuaf*.

Le mot zouave illustre parfaitement les deux cas de réalisation du [**و**]. Le TLFi, présente d'abord cette unité lexicale par une première indication phonétique du mot [zwa:v] accompagnée d'une attestation relevée dans le dictionnaire de l'académie depuis 1778.

[25] Le substantif *razzia* est transcrit *razia* comme mot vedette, dans le dictionnaire français-arabe de Ben Sedira.

Quant à la notice historique, elle signale la présence du mot dans la langue française depuis 1623, mais l'étymologie de cet emprunt est attribuée à l'arabe maghrébin [Zwāwa] en précisant la provenance de ce terme « nom d'une confédération de tribus kabyles de la région du Djurdjura, en Algérie, où étaient traditionnellement recrutés des soldats ».

5. La semi-voyelle prépalatale [ي] se réalise par [**i**] dans : caïd : **قايد**, *qaïd* ; kabyle, **قبايلى**, *qebaïli* ; zouidja, **زويجة**, *zouïja* ; zaouïa, **زاوية**, *zaouïa*, chéchia : *tacacit* ; razzia : **غازية**, *r'azia*.

Le TLFi présente d'abord la prononciation de l'emprunt razzia [Ra(d)zja], suivi par une attestation d'usage. Les différentes orthographes du mot sont signalées par ordre d'apparition : razia, gaze, gazia, razzia et au pluriel des razzias. La formation du pluriel est un indice d'intégration du mot dans la langue française. Par ailleurs, le TLFi attribue l'étymologie de l'emprunt razzia à l'arabe maghrébin [ġāziya] et à l'arabe classique [ġazwa] avec la valeur sémantique d'« expédition militaire, campagne; incursion, attaque ». Toutefois, le TLFi ne précise pas de quel arabe il s'agit. Par contre, le Petit Robert (2003 : 2182) et le dictionnaire des mots d'origine étrangère (1998 : 187) précisent que le mot razzia est un emprunt fait à l'arabe algérien.

6. *La dentale occlusive sourde emphatique* [**ط**] se réalise généralement par [**t**] dans toubib : **طبيب**, tebib ; coton : *qten* ; matelas : *matraḥ*. Cet emprunt est attribué par le TLFi à l'italien *materasso*, mais le *Petit Robert* et le *dictionnaire des mots d'origine étrangère* font remonter l'étymologie à l'arabe *matrah* « chose jetée à terre ».

3 AU NIVEAU MORPHOSYNTAXE

L'arabe dialectal, le kabyle et le français, sont des langues qui possèdent les genres masculin et féminin. Cela suscite quelques réflexions sur le changement de genre, la formation du nombre et, éventuellement, quelques remarques sur les modifications subies par les unités lexicales empruntées. Cela nous permettra de tirer quelques conclusions sur les influences exercées par la structure morphosyntaxique de la langue emprunteuse dans l'intégration des emprunts.

3.1 Changement de genre

Le corpus des mots français empruntés à l'arabe, que nous avons recueilli dans les deux dictionnaires bilingues, nous montre que certains emprunts à l'arabe, pour être intégrés, s'adaptent au système morphosyntaxique du français en changeant de genre. Ainsi, les emprunts suivants qui sont féminins en arabe deviennent masculins lors de leur passage en français : café : فهوة, qahoua ; calife : خليفة, khalifa ; couffin : قفة, qoffa.

Intéressons-nous maintenant aux mots arabes empruntés au français. En arabe, le masculin et le féminin se distinguent, dans l'ensemble, formellement au singulier. D'une manière générale, la formation du féminin arabe se réalise par l'adjonction d'un suffixe, morphème du féminin /a(*t*)/ pour un substantif ou un participe. Quant au kabyle, il forme généralement le féminin sur le masculin par préfixation et suffixation du morphème /t/.

Lors de son passage du français à l'arabe, l'emprunt ne change pas toujours de genre. Néanmoins, certains d'entre eux deviennent féminins par adjonction du /a/ final qui est la marque du féminin arabe. Ainsi, carrosse devient كرّوصة karrouça ; parasol devient وردة سول, ouarda-sol (subst. fém.) ; mètre devient ميترة, mitra (subst. fém.) ; kilomètre devient كيلومطرة, kilomètra (subst. fém.). Ce dernier emprunt est composé de kilo et de mètre, en arabe le substantif kilo est masculin. Il est possible que le mot mètre emprunté beaucoup plutôt au grec ait exercé une influence sur l'emprunt kilomètre emprunté plus tard par l'arabe. D'autre part,

des syntagmes entiers ont été admis en français tel que : salamalec (subst. masc.) mot composé en arabe de salam et de álaik.

Certains emprunts kabyles faits au français changent également de genre. Ils prennent les marques du féminin kabyle /t/ initial et final. Ainsi, baquet devient tabaqit (subst. fém.) ; carrosse devient takarrust (subst. fém.). D'autres, se forment par l'ajout du /a/ final comme dans damier qui devient damma (subst. fém.) ; kilomètre qui devient kilomitra (subst. fém.). Par ailleurs, certains emprunts, féminins en français, prennent la marque du masculin kabyle. Ainsi, école devient lakul ; écurie devient kuri(l). Ces emprunts au français sont en effet des syntagmes nominaux définis, commençant par le déterminant défini (l') pour l'école et l'écurie. Ces deux syntagmes sont admis entièrement dans le système morphosyntaxique du kabyle pour ne former qu'un seul mot. Nous émettons l'hypothèse que le mot écurie a subi l'influence de son équivalent kabyle adainin (subst. masc.) -qui désigne aussi le lieu destiné à loger les chevaux et autres équidés- et finit par prendre son genre. Quant à l'emprunt lakul, il est formé sur le même mode que lazuq, left, luz etc, tous substantifs masculins. Le mot cartouche (de fusil) devient akartuc (subst. masc.), mais ce terme a une histoire particulière. Le TLFi, note dans sa notice historique que le terme est attesté dans la langue française comme substantif féminin dès 1591. Cependant, ce dernier est un emprunt fait à l'italien cartoccio (subst. masc.) avec changement de genre « comme terme d'art militaire et de pyrotechnie ». Le kabyle n'a-t-il pas emprunté ce mot directement à l'italien ? En outre le TLFi, précise que le mot italien cartuccia, (subst. fém.) au sens de « petit morceau de papier » est un emprunt sémantique fait au français au cours du XIXe siècle cette fois au sens « d'art militaire ». Quoi qu'il en soit le mot cartouche est un substantif masculin en kabyle et désigne l'objet contenant la charge d'une arme à feu. Si les locuteurs kabylophones utilisent les emprunts français en les adaptant aux règles morphosyntaxiques de leur langue, le changement de genre est un indice d'intégration de l'emprunt dans la langue d'accueil.

Il est noter que les deux dictionnaires bilingues que nous avons consultés ne mentionnent pas le genre des lexèmes empruntés. Nous savons tous que ces outils s'adressent en particulier aux français qui sont amenés, de part leurs relations quotidiennes, à être en contact avec les indigènes. Or le changement de genre des mots comme *mètre*, *parasol*, cartouche, école, écurie etc, met l'utilisateur français en difficulté, car ces termes ont un genre différent dans sa langue. L'utilisateur n'a donc aucun autre moyen que l'usage pour opérer cette distinction.

3.2 Formation de nombre

La formation du pluriel en arabe dialectal se construit à travers des schèmes variés. Cependant, la quasi-totalité des emprunts au français que nous avons relevés obéissent à la règle de formation de pluriel externe à suffixe. Ce dernier s'obtient par adjonction du suffixe /at/ au substantif singulier finissant le plus souvent par la syllabe ouverte /a/ ou par a+ consonne. Ainsi, fourchette : ferchita > ferchitat ; litre : litra > litrat ; tasse : taça > taçat ; officier : fecian > fecianat ; sergent : serjan > serjanat ; hectare : qt'ar> qt'arat.

Il est à signaler que le *Dictionnaire français-arabe de la langue parlée en Algérie* ne note pas tous les pluriels des emprunts qui peuvent se former avec le suffixe /at/. Nous pouvons citer à titre d'exemple : journal, journal(at ?) ; hôtel, loutil(at ?) ; soc, sekka(at ?).

La langue kabyle distingue trois types de formation du pluriel : le pluriel interne (par alternance de voyelles interne), le pluriel mixte (par suffixation et par alternance vocalique et/ou consonantique) et le pluriel externe (par suffixation). L'ensemble des emprunts que nous avons recensés s'obtient par le pluriel externe. Ce dernier est dit régulier -parce qu'il n'opère

pas de modification à l'intérieur du mot- se forme par le suffixe /i-en/ (le i- étant la transformation de la voyelle initiale a), /ti-in/ (le i est une transformation du a initial et le t est la marque du féminin kabyle) et /at/. Ainsi, bataillon : abatiun > ibataiunen ; bidon, abidun > ibidunen ; bocal, abuqal > ibuqalen ; capitaine, aqabṭan > iqabṭanen ; cartouche, akartuc > ikartucen. Baquet, tabaqit > tibaqiin ; blouse, tabluzt > tibluzin ; brouette, tabruit > tibruidin ; carrosse, takarrust > tikarrusin ; machine, tamacint > timacinin. Par ailleurs, les emprunts à suffixe /at/ sont nombreux et font état d'un usage commun entre l'arabe dialectal et le kabyle. C'est le cas des mots comme : cantine, kantina > kantinat ; caporal, kabran > kabranat ; école, lakul > lakulat ; kilomètre, kilomitra > kilomitrat ; officier, fitian > fitianat ; poste, bosta > bostat.

L'attribution des caractéristiques du pluriel kabyle est un indice qui nous renseigne sur le degré d'intégration des emprunts français dans le système morphosyntaxique de la langue d'accueil. Certains emprunts vont jusqu'à prendre deux marques de pluriel : machine peut donner en kabyle koninir > ikoniniren ou ikoninirat. D'autres changent de nombre comme : les allumettes, zalamit (singulier).

4 AU NIVEAU SÉMANTIQUE

Certains emprunts sont admis dans l'arabe dialectal et le kabyle avec le même sens que nous leur connaissons en français. D'autres, par contre, subissent des modifications sémantiques. Nous nous pencherons sur les phénomènes de restriction sémantique, d'élargissement sémantique, de déplacement sémantique et de glissement conatif.

4.1 La restriction sémantique

Le mot français consigner a perdu tout autre signification en kabyle pour donner, dans le dictionnaire français-kabyle, akonsini avec le sens de défense de sortir. Le TLFi nous fournit les différents acceptions du terme consigner qui est d'abord entré dans la langue française avec le sens « *déposer une somme en garantie* », ensuite pour prendre le sens de « *maintenir prisonnier* », puis le fait de « *faire enregistrer des marchandises sur les livres des messagers, transporteurs* », enfin, plus récemment pour « *mettre des bagages à la consigne d'une gare* ». Nous voyons bien que dans un contexte de colonisation le mot *consigner* a subi une restriction de sens pour ne garder que le sens de « *maintenir prisonnier* ». En effet, pendant les opérations de l'armée française dans les douars, les militaires donnent la consigne au chef de village de défendre aux indigènes de quitter leur domicile.

La même restriction sémantique est observée dans le terme timbré qui donne en arabe dialectal algérien tenbri pour ne désigner que le « Papier destiné à la rédaction d'actes civils ou judiciaires soumis au droit de timbre et portant la marque de l'État, représentant le paiement de la taxe exigée ». Le TLFi nous offre d'autres sens de l'adjectif timbré qui « se dit de quelqu'un qui a le cerveau dérangé » ou « se dit d'une voix qui résonne bien ». L'emprunt timbré a donc subi une restriction sémantique, pour ne désigner que la nouvelle réalité coloniale, qui répond au besoin de la justice et de l'administration des indigènes.

4.2 L'élargissement sémantique

Le mot *amin* -lui-même emprunté à l'arabe au sens de *fidèle*, *honnête*, *loyal*- est défini dans le dictionnaire français-kabyle par « *maire de village kabyle* », autrement dit chef d'un douar si l'on veut être au plus près de l'organisation traditionnelle de la société kabyle. Le TLF*i* définit le mot *amin* comme « *Nom donné en Kabylie à un magistrat qui remplit des fonctions municipales, civiles, judiciaires* ». Le sens de *amin* au pluriel (*les amins)* est étendu par la

suite à tous les « *présidents de corporations d'artisans et de négociants en Afrique du Nord* ». Il « *désigne également des administrateurs de douanes* ». Nous constatons que la fonction de l'*amin* ne se borne plus à son rôle traditionnel mais elle s'étend pour embrasser d'autres fonctions telles que négociant, juge, douanier et, pendant la période coloniale, celui qui accompagne les militaires français dans les villages : « *Outre nos cavaliers d'escorte et nos gens de service, nous emmenons trois* ***amins****...* ». Nous voyons donc que le sens du mot *amin* s'est enrichi en français par extension sémantique de son acception kabyle.

L'extension sémantique s'observe aussi pour le mot *goum* qui signifie en arabe dialectal *tribu*, *peuple*, *gens*. Durant la colonisation française en Algérie, le mot *goum* est attribué au « *Contingent de combattants recrutés parmi la population indigène* ». Le TLF*i* note un dérivé du mot *goum* qui donne *goumier* au sens de « *Militaire faisant partie d'un goum* » et il s'étend par métonymie aux vêtements portés par ces soldats « *les manteaux blancs des* ***goumiers*** ».

4.3 Déplacement sémantique

Le mot romaine qui vient de l'arabe [rummān(a)] a tout à fait changé de sens en français. En effet, le mot romaine en arabe désigne « grenade », fruit du grenadier et par analogie il désigne le « poids-curseur de la romaine ». Le TLFi nous fournit la définition suivante : « Balance composée d'un fléau qui oscille autour d'un anneau d'attache le divisant en deux bras inégaux, le plus court muni d'un crochet où l'on suspend l'objet à peser, le plus long gradué et muni d'un curseur à poids invariable que l'on déplace pour atteindre l'équilibre du fléau à l'horizontale ». Le mot romaine signifie également « Laitue à feuilles allongées, fortement nervurées, croquantes ». Par ailleurs, TLFi note qu'un rapprochement avec le premier sens n'est peut-être pas à exclure. Nous constatons donc que le mot romaine d'origine arabe ne véhicule plus le sens de fruit de grenadier et par déplacement de sens, il évoque celui d'une balance et par extension peut-être celui d'une plante potagère.

4.4 Connotation négative

Le substantif salamalec emprunté à l'arabe [salāmalayk] qui signifie littéralement « paix sur toi », prend une connotation négative en français pour désigner « Politesses exagérées, révérences profondes ». Le TLFi recense quelques phrases où le mot salamalec a gardé une connotation négative dans des expressions telles que : « sans dire autre chose que leur **salem alec** ou Dieu vous garde », « Après avoir [...] dit par trois fois **salamalec** », « en faisant un nombre infini de **salamalec** ». Le TLFi note que le substantif masculin salamalec est d'un usage vieilli et qu'il s'utilise plutôt au pluriel dans un registre de langue familier. De la même façon, le mot caïd qui signifie en Afrique du nord « Notable qui cumule des fonctions administratives, judiciaires, financières; chef de tribu » prend une connotation négative pour qualifier un « mauvais garçon, chef de bande ». Le TLFi atteste cet emploi négatif dans des phrases de type : « Se prendre pour un caïd » et « Faire son caïd ». L'autorité du caïd, pendant la colonisation, inspire du respect et parfois de la crainte conférée par son pouvoir administratif. De plus, le caïd se révèle d'un caractère intraitable quant il s'agit d'appliquer une décision. C'est peut être cette dernière caractéristique qui a contribué à donner au mot caïd une connotation négative.

5 CONCLUSION

L'étude des emprunts recueillis dans les dictionnaires bilingues nous montre que pour s'intégrer dans la langue d'accueil, ces lexèmes subissent des modifications à différents niveaux. Ces changements sont dictés par le système linguistique de la langue d'accueil et

interviennent au niveau phonétique, phonologique, morphosyntaxique et sémantique. Or, le traitement lexicographique de l'emprunt dans les dictionnaires bilingues de la période coloniale est insuffisant. Nous avons vu à travers les deux ouvrages bilingues que nous avons utilisés que l'emprunt est traité d'une manière sommaire et lacunaire. Nous avons constaté que le changement de genre, par exemple, des lexèmes empruntés peut dérouter bien des usagers car les dictionnaires bilingues ne signalent pas ces changements. Si l'emprunt intervient pour garantir une communication sans équivoque entre les européens et les indigènes, le dictionnaire bilingue, outil consulté par les deux communautés, doit réserver un traitement satisfaisant aux unités empruntées. Le but du dictionnaire bilingue est de simplifier la communication interlinguale. Il gagne donc à être amélioré par l'indication des modifications phonétiques, phonologiques, morphosyntaxiques et sémantiques des emprunts. Signaler à l'usager d'un dictionnaire bilingue que l'emprunt appartient à une langue qu'il pratique, lui permet d'une part, de faciliter l'appropriation du terme et d'autre part, de prendre conscience des modifications subies par le mot lors de son intégration dans la langue d'accueil. Les emprunts de la période coloniale constituent, au-delà de la dimension culturelle, dans les dictionnaires bilingues, un cas d'observation de rapport de force entre les deux communautés. Ainsi, le dictionnaire bilingue peut offrir au grand public le moyen d'être informé d'une manière mesurée non seulement sur l'origine du mot mais aussi sur toutes autres informations linguistiques pertinentes et nécessaires. Pour les spécialistes, le TLF*i* fournit des informations précieuses sur l'origine de l'emprunt, le sens qu'il avait dans cette langue, des précisions sur la langue intermédiaire, etc.

Quant aux dictionnaires de la période coloniale, qui constituent un patrimoine linguistique, culturel et une source précieuse pour les études linguistiques, ils gagneraient à être complétés et actualisés par une version informatisée afin de préserver un lexique en langues autochtones qui tend à disparaître et de fournir aux futures générations une base de travail de référence.

6 RÉFÉRENCES

Asselah Rahal S. (2004). *Plurilinguisme et migration.* Paris : L'Harmattan. 262 p.

Ben Sedira B. K. (1886). *Dictionnaire français-arabe de la langue parlée en Algérie.* Alger : éd. Jourdan, 4 éd.

Boccuzzi C. (2007). « Traitement de l'anglicisme dans un corpus de dictionnaires bilingues ». *L'architecture du dictionnaire bilingue et le métier du lexicographe*. *A*ctes des Journées italiennes des dictionnaires, Capitolo-Monopoli, 16-17 avril 2007, sous la direction de Giovanni Dotoli, schena, Fasano (Italie). p. 155-164.

Cherifi N. (2008). *Les dictionnaires bilingues français-arabes et arabes-français : de leur traitement lexicographique et dictionnairique.* Thèse de Doctorat en Sciences du Langage, Université de Cergy-Pontoise. 485 p.

Dubois J. et al. (2002). *Dictionnaire de linguistique*. Paris : Larousse. 514 p.

Huyghe G. et Père Blancs (1902-1903). *Dictionnaire français-kabyle, Qamus Rumi-Qbaili.* France : L. & A. Godenne. 893 p.

Iraqui Sinaceur Z. (2004). « Histoire et emprunt linguistique ». Dans J. Daklia (dir.), *Trames de langues : usages et métissages linguistiques dans l'histoire du Maghreb.* Paris : Maisonneuve et Larose, coll. Connaissance du Maghreb. p. 509-524.

Laroussi F. (dir.) (1996). *Linguistique et anthropologie.* Cahiers de linguistique sociale. Université de Rouen. 1 156 p.

Lentin J. et Lonnet A. (2003). *Mélanges David Cohen : études sur le langage, les langues, les dialectes, les littératures, offertes par ses élèves, ses collègues, ses amis présentés à l'occasion de son quatre-vingtième anniversaire.* Paris : Maisonneuve et Larose. 770 p.

Nait-Zerrad K. (2001). *Grammaire moderne du kabyle.* Paris : Karthala. 225 p.

Queffelec A. et al. (2002). *Le français en algérie : lexique et dynamique des langues.* Bruxelles, Duculot, AUF, coll. Champs linguistiques. 590 p.

Quinsat F. (2008). « Le traitement lexicographique des arabismes dans les dictionnaires du français ». Dans J. F. Sablayrolles (dir.), *Néologie et terminologie dans les dictionnaires*. Paris : Honoré Champion. 241 p.

Reig D. (2006). *Dictionnaire arabe-français*. Paris : Larousse.

Samb A. et Puech M. (1978). *Grammaire arabe*. Dakar : Nouvelles éditions africaines. 187 p.

Sfar I. et Massoussi T. (2007). « Analyse et traitement lexicographique des emprunts ». Dans *Dictionnaires et mots voyageurs : l*es *40 ans du Petit Robert : de Paul Robert à Alain Rey*, Les Journées des dictionnaires de Cergy, sous la direction de Jean Pruvost. Eragny-sur-Oise : éd. des Silves. p.79-102.

Walter H. et Walter G. (1998). *Dictionnaire des mots d'origine étrangère*. Paris : Larousse, coll. Expression, 427 p.

LES REQUÊTES SUR UN SITE WEB : UN CORPUS POUR ÉTUDIER LA VARIATION ORTHOGRAPHIQUE

Jean-Luc Manguin
Laboratoire GREYC (UMR 6072 CNRS) – Université de Caen

RÉSUMÉ

Cette étude s'organise autour de deux idées ; tout d'abord, nous montrerons qu'un corpus de requêtes sur un site Web peut constituer le support d'un travail sur la variation graphique, au même titre qu'un corpus textuel tiré de textes en ligne. Pour cela nous nous appuierons sur des résultats statistiques comparant les variations graphiques observées dans les deux corpus. A la suite de ce premier résultat, nous examinerons les variations dans la transcription d'un motif particulier du français (le "double n") et nous verrons que le nombre d'erreurs dépend de la complexité du mot à transcrire, mais pas de sa fréquence ni de sa taille. En outre, nous serons à même d'interpréter certaines variations comme un phénomène de "brouillage" par des motifs graphiques concurrents qui apparaissent dans d'autres formes lexicales.

1 INTRODUCTION

L'arrivée des nouvelles technologies de l'information et de la communication (Internet, téléphonie mobile en mode texte) a fortement accru la production écrite, et en même temps diversifié les modes d'écriture. Ceux-ci varient en effet suivant la destination de cette production, qui peut entre autres être vouée à la lecture (sites en ligne, blogs), au dialogue rapide (chat, sms) ou encore à la communication homme-machine. Ce dernier type de communication inclut notamment les requêtes envoyées vers des ressources en ligne, sous forme de mots ou de groupes de mots délivrés de toute syntaxe. Cette quasi-absence de syntaxe se double néanmoins d'un effort de correction graphique, ceci afin d'éviter un bruitage de la réponse fournie par une machine destinataire généralement incapable de corriger les erreurs d'orthographe. Ce type de production textuelle peut donc a priori servir à l'étude de la variation graphique. Néanmoins, les mots observés étant dépourvus de tout co-texte, il importe de savoir si cet isolement influe ou non sur la quantité d'erreurs orthographiques observées. Pour répondre à cette question, nous allons démontrer que les observations sur un corpus de requêtes sont en accord avec celles que l'on peut faire sur des textes issus de forums de discussion. Dans une seconde partie, nous nous attacherons à un type d'erreur précis, et nous donnerons des explications sur les causes probables d'apparition de ce type d'erreur d'un point de vue général, complétées par des remarques sur certaines unités lexicales particulières. Ces explications feront appel à des résultats fournis récemment par la psycholinguistique.

2 LES DONNÉES OBSERVÉES

Le corpus d'observation est constitué des requêtes reçues par le dictionnaire des synonymes du Crisco [1998] pendant les neuf premières années de sa mise en ligne (oct. 1998 - déc.

2007) ; il représente plus de 200 millions de mots, répartis en un peu plus de 4 millions de formes distinctes. Nous n'avons retenu pour notre étude que les requêtes en mots simples, ce qui réduit la liste à 188 millions de mots pour 3,6 millions de formes différentes. Ensuite, pour examiner les formes et détecter parmi elles celles qui ne sont pas admises, nous avons choisi de nous restreindre aux 60500 formes qui ont été demandées 200 fois ou plus par les internautes. Nous avons pu alors, à l'aide de la base Morphalou, séparer les formes orthographiques admises de leurs variantes habituellement non-admises, ce qui nous a donné un premier fichier des "erreurs" orthographiques.

Dans un second temps, nous avons opéré une double sélection sur ces formes erronées : d'une part nous n'avons pas retenu celles ne comportant que des changements d'accents ou des omissions d'apostrophes ou de tirets, et d'autre part nous n'avons conservé que les "voisins orthographiques" des formes présentes dans la base Morphalou. Rappelons que deux mots sont des voisins orthographiques s'ils ne diffèrent que d'une lettre (ajout, suppression ou substitution). Le but de cette sélection était d'éviter la présence de difficultés difficiles à analyser, et d'erreurs purement dactylographiques (inversion de deux lettres, par exemple).

Nous avons ensuite classé les 4100 formes restantes suivant l'erreur commise, par ordre décroissant. Le tableau suivant en donne les 15 premières, avec leur fréquences respectives (et qui représentent 50 % du total des erreurs) :

Type de faute	**Fréquence**
double n → simple n	281
omission du e muet final	231
double r → simple r	181
double l → simple l	173
omission du s final	145
double m → simple m	139
double t → simple t	134
e → a , dans "en" ou "em"	127
double p → simple p	121
double s → simple s	113
ajout d'un e muet final	104
e → a , dans "an" ou "am"	95
simple l → double l	88
omission du h muet	83
simple n → double n	82

Tableau 1 : Les principales erreurs

On voit ici que parmi les 150 erreurs recensées, la famille "transformation d'une consonne double en consonne simple" monopolise les premières places dans le classement suivant la fréquence. Et si l'on va plus loin, l'erreur la plus fréquente, qui concerne le double n, se place

principalement sur le motif "onn[voyelle]" (220 cas sur 281). C'est la raison pour laquelle nous avons choisi d'observer les variations orthographiques sur ce motif.

Nos observations ont été faites sur un ensemble 351 formes, demandées plus de 1000 fois dans leur version correcte, et dont la version avec un simple n a été elle aussi demandée en nombre suffisant pour l'inclure dans une étude statistique. A titre d'illustration, nous donnons dans le tableau 2 les 20 formes les plus fréquentes, avec la proportion d'erreurs relevée dans notre corpus.

Référence	**Nb1**	**Erreur**	**Nb2**	**Proportion**
donner	161654	doner	880	0,54%
connaissance	100474	conaissance	1404	1,40%
personne	61020	persone	539	0,88%
connaitre	56401	conaitre	827	1,47%
connaître	47264	conaître	251	0,53%
exceptionnel	44216	exceptionel	1778	4,02%
reconnaissance	40965	reconaissance	747	1,82%
bonne	39987	Bone	532	1,33%
professionnel	39242	professionel	6215	15,84%
fonctionnement	39208	fonctionement	667	1,70%
environnement	38772	environement	2696	6,95%
abandonner	38655	abandoner	768	1,99%
mentionner	38367	mentioner	1143	2,98%
personnel	35551	personel	2112	5,94%
passionné	33523	passioné	5589	16,67%
inconnu	27209	inconu	472	1,73%
personnage	27024	personage	447	1,65%
fonctionner	26806	fonctioner	393	1,47%
connu	24530	Conu	126	0,51%
raisonnable	23892	raisonable	2590	10,84%

Tableau 2 : Les 20 formes les plus fréquentes en "onn[voyelle]"

3 LA COMPARAISON AVEC UN AUTRE CORPUS

Comme nous l'avons dit en introduction, la validation de nos observations sur notre ensemble de requêtes passe par une comparaison avec un corpus textuel où les mots sont pourvus d'un contexte (ou, si l'on préfère, d'un co-texte). Cependant, il est exclu d'employer pour cette comparaison un corpus qui serait passé par le filtre du correcteur orthographique ou par celui de la relecture. Nous devons donc choisir un corpus dont les conditions de dactylographie se rapprochent de celles où l'on tape une requête dans un formulaire en ligne. Mais en même temps, ce corpus ne devra pas être spécialisé, afin d'éviter certains biais dans les fréquences

observées (par exemple, le mot "abonnement" est quasiment toujours orthographié correctement sur les forums relatifs à Internet).

Toutes ces raisons font que nous avons finalement opté pour un ensemble de forums en ligne, que nous avons interrogés manuellement par Google®, car l'interface de ce moteur de recherche permet de faire des requêtes sur plusieurs sites à la fois, grâce à l'emploi de caractères "joker". Ainsi, nous avons cherché nos formes correctes et erronées sur "forum.*.fr", ce qui se traduit par des recherches sur les sites suivants (entre autres) :

30. forum.gulli.fr
31. forum.tf1.fr
32. forum.doctissimo.fr
33. forum.ados.fr
34. forum.elle.fr
35. forum.sfr.fr
36. forum.europe1.fr
37. forum.hardware.fr
38. forum.automoto.fr
39. forum.letudiant.fr

On voit, au travers des noms de domaines récupérés par cette méthode, que l'ensemble constitue un éventail assez varié, ce qui, en principe, doit nous mettre à l'abri des biais mentionnés plus haut. Nous avons donc relevé sur cet ensemble les fréquences des graphies normales et erronées de nos 351 formes précédemment évoquées, puis calculé le pourcentage d'erreurs pour chacune de ces formes. Le tableau 3 donne les proportions dans les deux corpus, pour les 20 formes du tableau 2.

Forme admise	Corpus de requêtes	Corpus de forums
donner	0,54%	0,54%
connaissance	1,40%	1,55%
personne	0,88%	1,34%
connaitre	1,47%	1,38%
connaître	0,53%	0,11%
exceptionnel	4,02%	3,56%
reconnaissance	1,82%	2,78%
bonne	1,33%	1,26%
professionnel	15,84%	10,67%
fonctionnement	1,70%	0,25%
environnement	6,95%	0,65%
abandonner	1,99%	1,67%
mentionner	2,98%	2,22%
personnel	5,94%	0,57%
passionné	16,67%	22,86%
inconnu	1,73%	1,18%
personnage	1,65%	2,89%
fonctionner	1,47%	0,95%
connu	0,51%	0,30%
raisonnable	10,84%	9,94%

Tableau 3 : Proportion de formes erronées dans les deux corpus

Il est facile de remarquer dans ce tableau la similarité des proportions d'erreurs, sauf pour "environnement" et "passionné". Pour confirmer la similitude des observations dans les deux corpus, nous avons fait une analyse statistique, en calculant le coefficient de corrélation entre les deux séries de données. Notons toutefois qu'il s'agit ici de comparer des quotients, c'est la raison pour laquelle nous avons fait cette analyse sur les logarithmes décimaux des proportions d'erreurs ; nous pensons en effet que dans le cas qui nous préoccupe, c'est l'ordre de grandeur de la proportion qui est important, et non sa valeur absolue. En d'autres termes, si une forme présente 1% d'erreurs dans un corpus et 2% dans l'autre, la différence équivaut au cas d'une autre forme qui montrerait 10% dans un corpus et 20% dans l'autre. Signalons enfin que le calcul avec les valeurs absolues des proportions corrobore celui fait avec les logarithmes. Le coefficient de corrélation trouvé entre les deux séries de logarithmes des proportions est égal à 0,76, ce qui montre un bon accord entre les deux corpus. Si l'on élimine les 10% qui correspondent aux plus gros écarts, le coefficient de corrélation atteint 0,87. Nous en donnons ci-dessous la courbe représentative (figure 1).

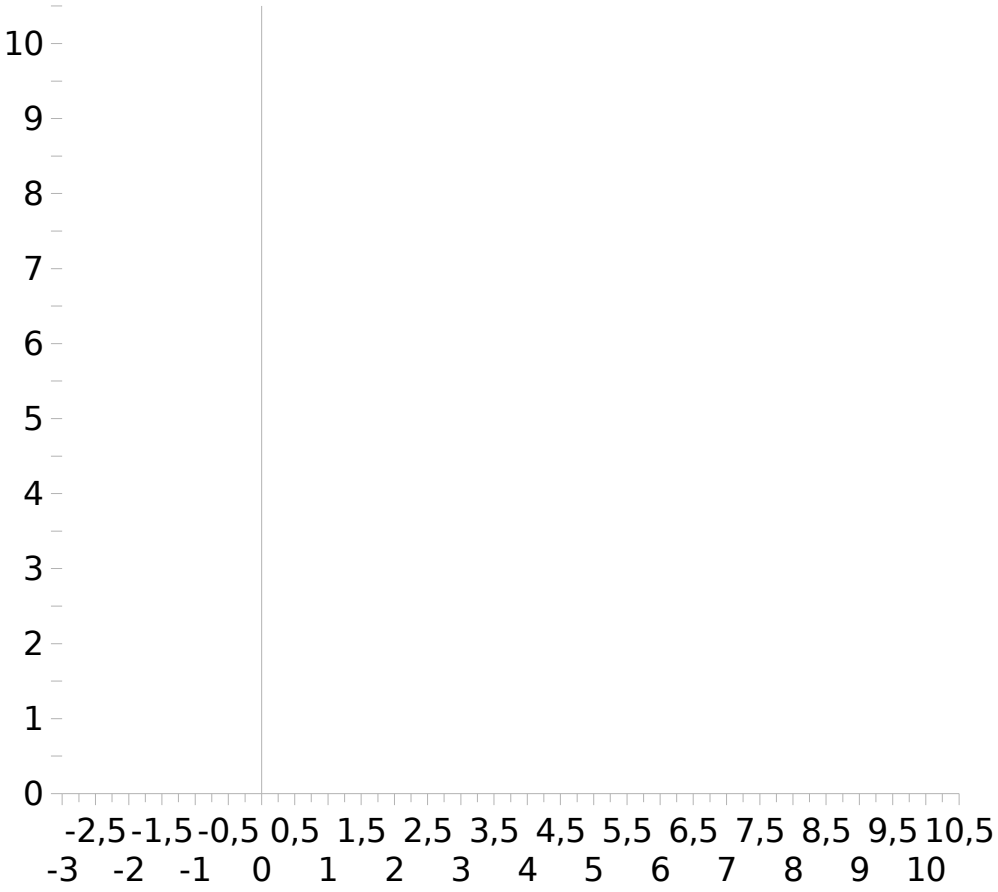


Figure 1 : Proportions d'erreurs dans l'un et l'autre corpus (échelle logarithmique)

On peut ainsi conclure cette première partie en disant que malgré des écarts individuels sur certaines formes, notre corpus de requêtes constitue un corpus valable pour l'observation de la variation graphique.

4 OBSERVATIONS GÉNÉRALES

Nous commençons cette étude des erreurs orthographiques par des observations générales, autrement dit statistiques, sur notre corpus. Dans un second temps, nous chercherons à expliquer certaines erreurs par des propriétés particulières du mot dans lequel se produit l'erreur.

Le premier paramètre à prendre en compte est la longueur des mots. Nous savons, grâce notamment aux travaux de Caramazza (cité par Bonin [2007]), que le taux d'erreurs orthographiques est normalement indépendant de la longueur des mots ; une dépendance serait en effet un des signes d'une atteinte cérébrale au niveau du buffer graphémique. En ce qui concerne nos données, nous n'avons pas trouvé de corrélation entre la longueur des mots et leur proportion d'erreurs (coefficient de corrélation égal à 0,1). Notre corpus est donc sur ce point en accord avec le modèle "normal" de production verbale écrite.

Nous n'avons pas trouvé non plus de corrélation entre la proportion de la forme erronée et la fréquence de la forme correcte, ni dans notre corpus, ni dans l'autre. Ce point peut porter à discussion ; il existe en effet de nombreux travaux concernant les relations entre les erreurs commises sur un mot, la fréquence de celui-ci et son âge d'acquisition. Malheureusement pour

nous, nous n'avons pas du tout d'information sur les auteurs des requêtes en ligne. Il est donc probable que toutes les différences individuelles aient été gommées par le mode de recueil global que constituent les fichiers de traces d'un serveur Web. Il pourrait toutefois être intéressant de tenir compte du domaine d'origine de la machine émettant la requête, bien que cette donnée ne soit pas d'une grande fiabilité.

Enfin la fréquence de certaines variations peut être examinée en faisant appel à la notion de "difficulté phonogrammique" introduite par N. Catach [2001, 2003] ou J.P. Jaffré [2008]. Les psycholinguistes utilisent le terme de difficulté "phono-graphique" pour parler aussi de ce genre de difficulté qui se rencontre lorsqu'il peut y avoir plusieurs manières de transcrire un son, comme c'est le cas en français par exemple pour le son "i", le son "s" ainsi que toutes les consonnes qui peuvent être doublées. On dit également qu'il y a dans ce cas un problème de "consistance", et plus précisément pour celui des consonnes doubles, une "inconsistance phono-graphémique" (c'est-à-dire : à un seul son correspondent plusieurs graphèmes, pour ces notions, voir par exemple Ferrand [2001]).

Nous avons donc examiné la proportion d'erreurs commises en fonction de la complexité phonogrammique d'un mot, autrement dit en fonction du nombre de difficultés de même ordre phonogrammique que doit résoudre le scripteur. En l'occurrence, un mot qui contient deux consonnes doubles sera plus complexe qu'un mot qui n'en contient qu'une. Nous avons ainsi étudié la fréquence des fautes sur le "double n", en fonction ou non de la présence d'une autre consonne géminée dans le mot. Si l'on répartit les 351 mots précédemment étudiés en deux groupes, ceux qui contiennent une autre consonne géminée, et ceux qui n'en contiennent pas, on obtient une différence significative : pour le groupe à deux consonnes, la proportion moyenne d'erreurs est 17,2 %, tandis qu'elle ne vaut que 6,9 % pour l'autre groupe (les données avec le corpus des forums donnent des proportions encore plus différenciées : 16,2 % et 4,7 %). Rappelons au passage qu'il n'est question ici **que des erreurs commises sur le double n** ; il va de soi que les mots présentant deux consonnes doubles - dont le double n - peuvent présenter deux autres formes erronées (avec erreur sur l'autre consonne, et avec erreurs sur les deux consonnes). Ces résultats sont résumés dans le graphique de la figure 2.

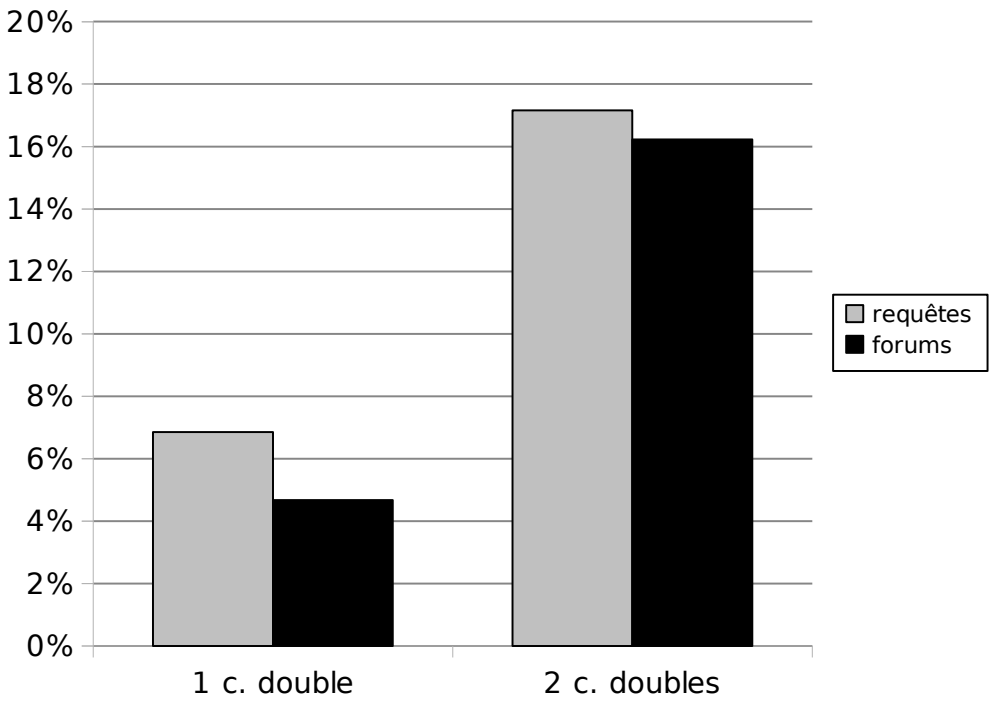


Figure 2 : Proportion d'erreurs sur le double n en fonction du nombre de consonnes doubles

Bien que ce résultat ne soit pas très surprenant, puisqu'il rappelle certaines observations faites chez des patients atteints de dysgraphie lexicale (voir par exemple Beauvois et Dérouesné, cités par Bonin [2007]), nous pensons poursuivre nos observations afin de vérifier si cette tendance est généralisable à toutes les consonnes doubles de la langue française.

Pour aller plus loin, nous avons divisé le groupe des formes contenant deux consonnes doubles en deux sous-groupes, suivant que le double n se situe avant ou après l'autre consonne double. Nous avons là encore trouvé une différence significative entre ces deux groupes : si le double n se trouve **avant**, la proportion d'erreurs n'est que de 13,3 %, tandis qu'elle passe à 19,8 % s'il se trouve **après** l'autre consonne double. Ce résultat est illustré par la figure 3.

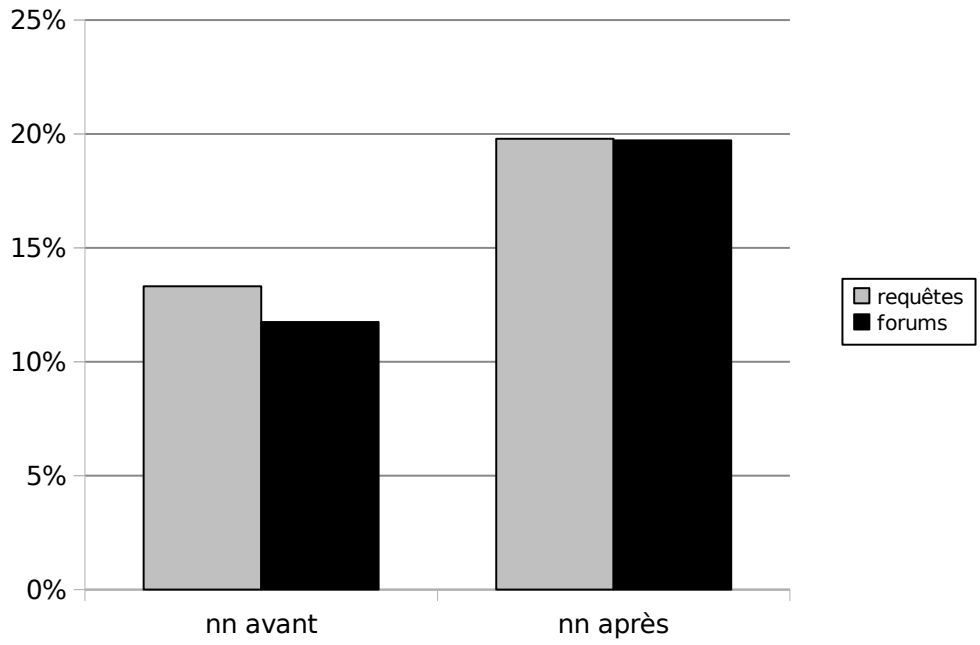


Figure 3 : Proportion d'erreurs suivant la position du double n, dans les mots à deux consonnes doubles

Ce résultat va dans le même sens que les observation faites sur la langue anglaise par Wing et Baddeley [2009] et qu'ils interprètent comme des effets de mémoire dans le buffer graphémique.

5 OBSERVATIONS PARTICULIÈRES

Au-delà des remarques générales que nous venons de formuler, signalons que si la valeur moyenne des erreurs est 9,8 %, des disparités existent cependant en fonction de la voyelle qui suit le double n. Ces différences sont résumées dans la figure 4 ci-après.

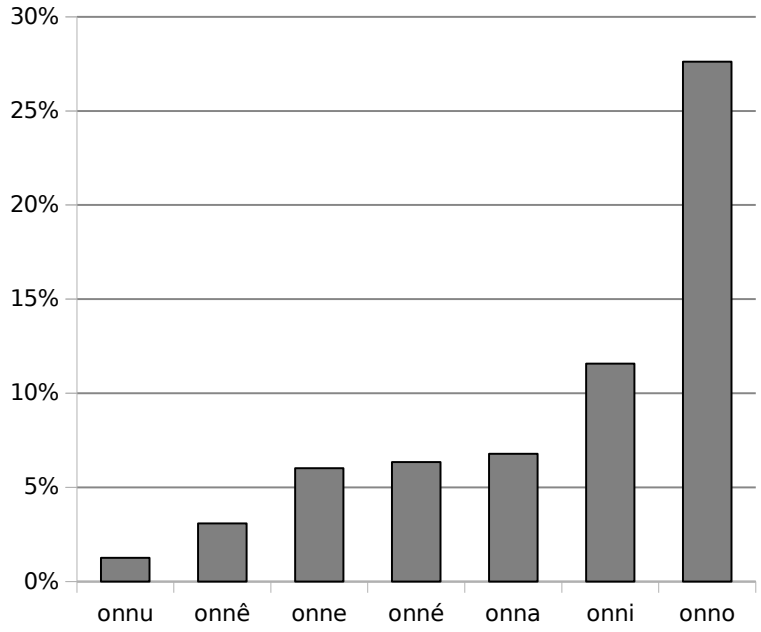


Figure 4 : Proportion d'erreurs en fonction de la voyelle suivante

Dans le cadre de cet article, nous nous pencherons sur les cas extrêmes, et nous donnerons des explications issues des travaux de psycholinguistique. Tout d'abord, le motif "**onnu**", qui possède le plus faible pourcentage d'erreurs (1,3 %), ne se rencontre que dans les formes conjuguées de "connaître" et ses dérivés (et dans notre corpus, nous n'avons affaire qu'aux participes passés). Ainsi nous avons déjà deux explications à l'écriture quasiment toujours correcte de ce motif :

7. Il est situé en fin de mot ; et selon Wing et Baddeley [2009], les extrémités des mots sont moins porteuses d'erreurs.
8. Le "u" (et ce qui peut le suivre) est lié à une marque morphologique ; les observations des neuropsychologues confortent en effet l'hypothèse d'un niveau de représentation qui code la structure morphologique des mots. Du coup, le double n se trouve à la fin d'une partie de la structure, ce qui renforce sans doute sa mise en mémoire.

Il y a en outre une troisième raison à la qualité de production de ce motif "onnu" : c'est la fréquence de ce motif par rapport à son concurrent "onu". Les travaux de S. Pacton (voir par exemple Danjon et Pacton [2009]) ont mis en évidence l'apprentissage des régularités graphotactiques de la langue et leur mémorisation. De plus, le motif "onu" n'est jamais une marque morphologique, et se trouve soit en milieu de mot comme dans "monument" ou en fin de mot avec un "s" final qui se prononce, comme dans "bonus" ou "tonus".

Par contre, pour "**onno**" (27,6 % d'erreurs), il convient tout d'abord d'annoncer que parmi nos 351 formes, seuls 3 mots contiennent ce motif : "entonnoir", "connotation" et "connoter".

Il semble donc plus judicieux d'élargir nos recherches dans notre corpus, ce qui nous conduit à distinguer plusieurs cas. Pour "onnons" qui se rencontre en fin de verbe conjugué, les fautes sont moins fréquentes (10,9 %), car le motif "onon" est rare en fin de mot (sauf dans des noms propres comme "Donon", "Thonon") et qu'il ne se trouve en milieu que dans "prononcer" et ses dérivés. Le tableau 4 ci-après donne les occurrences relevées dans notre corpus.

	Normal	erreurs	proportion
mentionnons	135	20	14,8 %
donnons	118	7	5,9 %
perfectionnons	31	4	12,9 %

Tableau 4 : Formes relevées avec le motif "onnons"

Pour "entonnoir" (16,4 % d'erreurs), on s'aperçoit à la lecture du tableau 5 que sa famille lexicale n'est pas affectée de la même manière par les erreurs orthographiques. On peut donc en déduire que la finale "noir" devient prépondérante, et qu'il existe un brouillage avec "manoir", "urinoir" et "laminoir", car "entonne" et "entonné" ont un pourcentage d'erreurs nettement plus faible.

	Normal	erreurs	proportion
entonnoir	2162	354	16,4 %
entonnoirs	19	5	26,3 %
entonner	2764	192	6,9 %
entonné	67	4	6,0 %

Tableau 5 : "entonnoir" et sa famille

Enfin, le cas de "connotation" (plus de 30 % d'erreurs pour "connoter" et ses dérivés) est plus complexe. Le tableau 6 donne les observations dans notre corpus.

	Normal	Erreurs	proportion
connotation	7160	1661	23,2 %
connotations	251	29	11,6 %
connoter	2581	1273	49,3 %
connoté	632	201	31,8 %
connotatif	286	98	34,3 %

Tableau 6 : "connotation" et sa famille

Certes la relative rareté d'emploi de ces mots et leur niveau de langue soutenu pourraient avoir une influence sur la quantité d'erreurs commises, mais nous penchons plutôt pour un faisceau convergent d'influences :

8. l'antonyme "dénotation" ne prend qu'un seul n.
9. la construction fréquente avec le préfixe "co-" sans doublement de la première consonne.
10. la présence très habituelle du motif "cono" dans "économie" et tous ses dérivés.

6 CONCLUSION ET PERSPECTIVES

Nous avons ainsi démontré dans la première partie de cet article que les nombreuses requêtes recueillies par les formulaires en ligne peuvent aussi servir à amasser du matériel présentant

un intérêt linguistique, notamment pour l'observation des variantes orthographiques. Ce mode de recueil est d'autant plus intéressant qu'il est facilité par l'inscription automatique des traces des internautes.

Concernant les erreurs orthographiques, nous avons montré que d'une manière générale la difficulté phonogrammique liée à la transcription du son "n" en double "n" peut varier en fonction de la connaissance lexicale d'un mot et de sa famille, mais subit avant tout l'influence de mécanismes inconscients de mémorisation et d'attention. Nous avons en effet relevé que la transcription d'un motif phonétique peut être perturbée par le report de l'attention vers une autre difficulté dans le mot à transcrire, mais aussi par l'habitude de transcrire ce motif phonétique selon un autre schéma graphique plus courant. Il est évident que ce travail constitue une première étape, et qu'il serait très intéressant de poursuivre de manière plus complète l'étude des mots à consonnes doubles et de leurs variantes produites.

7 RÉFÉRENCES

Bonin P. (2007). *Psychologie du langage*. De Boeck.

Catach N. (2003 nouvelle éd.). *L'orthographe*. PUF, collection "Que sais-je ?".

Catach N. (2001). *Histoire de l'orthographe française.* Honoré Champion.

CRISCO (1998). Dictionnaire électronique des synonymes. http://www.crisco.unicaen.fr/

Ferrand L. (2001). *Cognition et lecture*. De Boeck.

Jaffre J-P. (2008). *Nouvelles recherches en orthographe*. Lambert-Lucas.

MORPHALOU. Lexique morphologique ouvert du français. http://www.cnrtl.fr/lexiques/morphalou/

Danjou J. et Pacton S. (2009). « Apprentissages implicites dans l'acquisition de l'orthographe ». *Entretiens de Bichat*, 35-49.

Wing A. et Baddeley A. (2009). « Righting errors in writing errors: The Wing and Baddeley (1980) spelling error corpus revisited ». *Cognitive neuropsychology*, vol. 26, 2009, no 2, p. 223-226.

EXPLORER DES CORPUS À L'AIDE DE CASSYS. APPLICATION AU *CORPUS D'ORLÉANS*

Denis Maurel[1], Nathalie Friburger[1], Iris Eshkol[2] et Jean-Yves Antoine[1]

[1]Laboratoire d'informatique – Université François Rabelais Tours

[2]Laboratoire ligérien de linguistique – Université d'Orléans

1 INTRODUCTION

Cet article présente un outil d'exploration de corpus, *CasSys*, facilement paramétrisable par les linguistes, permettant de reconnaître des motifs même complexes et de les baliser, éventuellement par des balises XML. Ce balisage automatique peut ensuite être révisé par un expert. CasSys est donc un outil d'exploration de corpus, mais également d'annotation enrichie semi-supervisée.

Deux exemples réels complèteront cette présentation : la recherche des entités nommées du *Corpus d'Orléans* et l'utilisation de ces entités pour connaître des informations sur les personnes répondant à l'enquête qui constitue ce corpus. Ce travail a bénéficié du financement du projet ANR *Variling* et d'un projet Feder Région Centre. Il a aussi été testé dans le cadre de l'évaluation *Ester2* (*campagne d'évaluation des systèmes de transcription enrichie d'émissions radiophoniques*)[1].

2 LE SYSTÈME CASSYS

Le système CasSys [Friburger, 2002] [Friburger, Maurel, 2004] est un programme permettant le passage sur un corpus de transducteurs "en cascade" (c'est-à-dire les uns après les autres) dans un ordre défini [Abney, 1996]. Ces transducteurs sont des graphes Unitex [Paumier, 2003], facilement manipulables et modifiables par un linguiste grâce à une interface conviviale. Il est donc possible de définir des cascades pour différents balisages.

1. Annotation de corpus : moyennant quelques adaptations, CasSys peut s'adapter à tout corpus, nous l'avons appliqué au journal Le Monde, à Eslo1 et Ester2, pour la reconnaissance des entités nommées (cascade *CasEN*).
2. Exploration de corpus : par exemple, une fois les entités nommées balisée, la cascade *CasDen* reconnaît les entités dénommantes, c'est-à-dire les informations concernant la personne interrogée (composition familiale, profession, lien avec Orléans…).

La Figure 1 présente un exemple de transducteur de la cascade CasEN. Celui-ci reconnaît et balise les établissements de soin. Des transducteurs précédents ont déjà reconnu une personne (idxPerson), éventuellement avec un titre (idxTitre), ou une organisation (idxOrg), ou encore un lieu (idxLoc)… Par exemple, *hôpital Jean Rostand*, *clinique Docteur Calabet d'Agen*, etc. qui seraient placés entre les balises <ENT type=loc.fac> et </ENT>.

[1] http://www.afcp-parole.org/ester

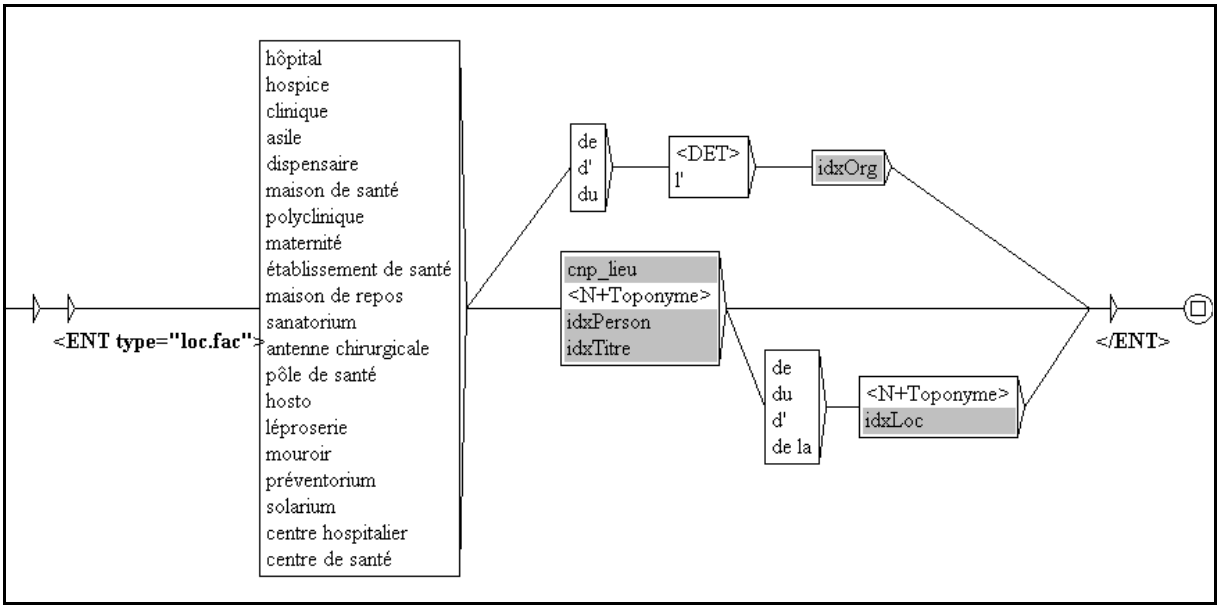


Figure 1 : Un exemple de transducteur

3 LE *CORPUS D'ORLÉANS*

Le corpus sur lequel nous avons travaillé est l'*Enquête sociolinguistique à Orléans* (*Eslo1*), plus précisément, les cent vingt premières heures d'interview transcrites (cent cinq fichiers *Transcriber*, soit 31 004 Ko). Nous comptons utiliser aussi bientôt le corpus *Eslo2*, en cours de constitution, en partant des acquis d'Eslo1, une nouvelle enquête a été mise en chantier par le LLL. Il s'agit, à quarante années de distance, de constituer un corpus comparable dans le produit attendu et dans les modalités de la collecte (400 heures, environ 6 000 000 de mots).

Les principales conventions de transcription sont l'absence de ponctuation et de majuscule en début d'énoncé ainsi qu'une transcription orthographique normée (majuscule pour les noms propres, transcription des chiffres et des dates en toutes lettres avec les traits d'union si nécessaire, termes épelés notés entièrement en majuscule). Le questionnaire de l'entretien contient tout d'abord des questions préliminaires (Depuis combien de temps habitez-vous Orléans ?, Qu'est-ce qui vous a amené à vivre à Orléans ?, Est-ce que vous vous plaisez à Orléans ?, etc.), puis des questions sur le travail et les loisirs du locuteur et des membres de sa famille, ce qui explique la présence d'un nombre important d'entités nommées dans le corpus.

4 LA CASCADE CASEN

La cascade CasEN a tout d'abord été conçue pour reconnaître les entités nommées de corpus journalistiques (essentiellement *Le Monde*) [Friburger, 2002]. En 2006, dans le cadre du projet ANR Variling, elle a été reprise et adaptée au corpus d'Orléans. De plus, le balisage des entités nommées a été modifié pour correspondre à celui de la campagne Ester 2, à quelques ajouts près. Cette typologie est présentée Figure 2.

personne (*pers*)	humain réel ou fictif (*pers.hum*) animal réel ou fictif (*pers.anim*)	
fonction (*fonc*)	politique (*fonc.pol*) militaire (*fonc.mil*) administrative (*fonc.admi*) religieuse (*fonc.rel*) aristocratique (*fonc.ari*)	

organisation (*org*)	politique (*org.pol*) éducative (*org.edu*) commerciale (*org.com*) non commerciale (*org.non-profit*) média & divertissement (*org.div*) géo-socio-administrative (*org.gsp*)	
lieu (*loc*)	géographique naturel (*loc.geo*) région administrative (*loc.admi*) axe de circulation (*loc.line*) construction humaine (*loc.fac*)	
	adresse (*loc.addr*)	adresse postale (*loc.addr.post*) téléphone et fax (*loc.addr.tel*) adresse électronique (*loc.addr.elec*)
production humaine (*prod*)	moyen de transport (*prod.vehicule*) récompense (*prod.award*) œuvre artistique (*prod.art*) production documentaire (*prod.doc*)	
date et heure (*time*)	date (*time.date*)	date absolue (*time.date.abs*) date relative (*time.date.rel*)
	heure (*time.hour*)	
montant (*amount*)	valeur physique (*amount.phy*)	âge (*amount.phy.age*) durée (*amount.phy.dur*) température (*amount. phy.temp*) longueur (*amount. phy.len*) surface et aire (*amount.phy.area*) volume (*amount.phy.vol*) poids (*amount. phy.wei*) vitesse (*amount. phy.spd*) autre (*amount.phy.other*)
	valeur monétaire (*amount.cur*)	

Figure 2 : La typologie Ester[2]

Cependant ces annotations ne nous ont pas paru suffisantes sur deux points. Tout d'abord, il nous a semblé intéressant d'annoter et de détailler l'ensemble des informations sur les personnes, lorsque celles-ci constituent une expansion classifiante de l'entité. Par exemple simplement sa civilité, mais aussi son origine géographique (gentilé, ethnique…), sa profession, sa religion, son appartenance politique, etc. De plus nous avons aussi ajouté un balisage des dynasties. Ensuite, nous avons aussi complété les différents types d'Ester par un typage des évènements, parmi lesquels nous avons distingué les faits historiques et les différentes manifestations sportives, culturelles, etc. La Figure 3 détaille ces ajouts.

pers.hum.tit	les civilités (M. Mme, Melle, Monsieur, Madame…)
pers.hum.gent	les gentilés (Tourangeau, Parisien…) les adjectifs toponymiques (tourangeau, parisien…)

[2] http://www.afcp-parole.org/ester/docs/Conventions_EN_ESTER2_v01.pdf

pers.hum.occ	les professions (avocat, journaliste…)
pers.hum.sp	les sports (coureur cycliste, footballeur…)
pers.hum.art	les artistes (violoniste, sculteur…)
pers.hum.nat	la nationalité (français, hollandais…)
pers.hum.rel	la religion (musulman, catholique…)
pers.hum.pol	la politique (communiste, socialiste…)
pers.hum.fonc	les titres professionnels (Professeur, Maitre, Docteur…)
pers.hum.dynasty	(les Bourbons, les Windsor…)
event	les évènements
event.hist	l'histoire (la Seconde Guerre mondiale, la Prise de la Bastille…)
event.manif	les manifestations sportives, artistiques (les Jeux olympiques, les Francofolies…)

Figure 3 : Complément aux annotations d'Ester

La cascade *casEN* commence par lancer le programme de préanalyse d'*Unitex*, en utilisant un graphe inspiré de celui de [Dister, 2007], puis le dictionnaire Delas [Courtois., Silberztein, 1990] et des dictionnaires spécifiques contenant 28 341 prénoms, 31 580 professions [Gazeau, Maurel, 2006], 3 016 sigles, 114 511 noms propres (et dérivés de noms propres) extraits de Prolexbase [Tran, Maurel, 2006], 497 noms d'animaux, 296 noms de sports, 110 noms de monnaies, 53 noms de marques de voiture et 26 noms de quotidiens. Elle est suivie de 152 transducteurs passés en cascade.

Voici quelques exemples de balisage :

il y a deux ans une euh <ENT type="pers.hum.nat"> anglaise </ENT>

chez moi <ENT type="pers.hum"> Bérénice Nutal </ENT>

dans les <ENT type="org.com"> PTT </ENT>

moi je suis native de <ENT type="loc.admi"> Pithiviers </ENT> j'aime mieux <ENT type="loc.admi"> Orléans </ENT>

<ENT type="pers.hum"> Sophie </ENT> viens voir

oh j'ai une <ENT typr="prod.art"> encyclopédie Quillé </ENT> j'ai le

<ENT type="time.date.abs"> en dix-neuf cent trente-huit </ENT>

je crois que le <ENT type="org.pol"> ministère de l'Education National </ENT>

le <ENT type="org.edu"> lycée Pothier </ENT> et les élèves qui vont au <ENT type="org.edu"> lycée Benjamin Franklin </ENT>

euh passer quelques jours sur la <ENT type="loc.geo"> Côtes d'Azur </ENT>

euh je suis je travaille à l'<ENT type="loc.fac"> hôpital d'Orléans </ENT> quoi

en <ENT type="loc.admi"> Norman- Normandie </ENT> peut-être ?

parce que nous avions un <ENT type="loc.fac"> magasin Phildar </ENT> juste en face de chez nous

oui c'est un petit <ENT typr="prod.art"> dictionnaire Larousse </ENT>

qui vont se charger au maximum quand le gars aura le dos tourné vous connaissez ben l'histoire de <ENT type="pers.hum"> Marius et Fanny </ENT> et tout ça

j'ai des copains qui y travaillent et c'est très intéressant ils ont fait le <ENT type="loc.fac"> pont de Tancarville </ENT> euh

dans la <ENT type="loc.line"> rue Royal</ENT> euh

L'ensemble du corpus Eslo1 a ainsi été annoté et ces annotations ont été vérifiées manuellement ensuite. L'évaluation des résultats du passage de la cascade CasEN a été détaillée dans [Maurel et al., 2009]. Un résumé de cette évaluation est présenté Figure 4. Puisqu'une relecture des entités trouvées était prévue dans le projet, il nous a paru intéressant de mesurer tout d'abord la simple détection des entités, en acceptant d'éventuelles erreurs de typage ou de bornage. CasEN a aussi été évaluée dans le cadre de la campagne Ester2 [Galliano et al., 2009].

Entités	Non typée	Partielle	Complète
Rappel	94,0%	88,4%	87,5%
Précision	97,8%	92,0%	91,1%

Figure 4 : Résultats de l'évaluation

5 LA CASCADE CASDEN

Le corpus Eslo1 était, à l'origine, une enquête sociolinguistique. Il nous a paru intéressant, dans le cadre du projet Variling, de proposer au lecteur des informations sur les personnes interviewées et sur leur famille. Nous avons appelé ces donnés *entités dénommantes*. Elles permettent de mieux connaître sociologiquement le locuteur. Peut-être certaines de ces informations seront d'ailleurs cachées dans le cadre de la distribution libre du corpus [Eshkol, 2007], afin de respecter l'anonymat des témoins [Baude, 2006]. Nous avons donc créé une nouvelle cascade, *CasDen*, pour justement repérer automatiquement ce genre d'information. Celle-ci passe non pas sur le texte originel, mais sur le texte balisé par la cascade CasEN.

Nous avons donc défini de nouveaux types pour décrire la personne qui parle ou dont on parle (identité, famille, travail, syndicat, âge, origine…). Ces types sont présentés Figure 5. Parfois nous avons ajouté des informations quantitatives (nombre d'enfants, âge…) dans les balises.

pers.speaker	la personne interviewée
pers.spouse	son époux ou épouse
pers.child	ses enfants
pers.parent	ses autres liens de parenté
identity.age	l'âge
identity.origin	l'origine géographique
identity.birth	la date de naissance
identity.arrival	la date d'arrivée à Orléans
identity.children	l'identité de ses enfants
work.occupation	le métier
work.field	le domaine professionnel
work.location	le lieu de travail
work.business	l'entreprise
trade union	l'appartenance syndicale

Figure 5 : Typologie des annotations de la cascade CasDen

Voici par exemple le traitement d'une question sur l'arrivée de l'interviewé à Orléans :

depuis combien de temps habitez vous <ENT type="loc.admi">Orléans</ENT> ?
<DE type="pers.speaker"><DE type="identity.origin">
<Turn speaker="spk1" startTime="6.754" endTime="10.88">
oh ça fait <ENT type="time.date.rel">neuf ans</ENT> depuis dix neuf cent soixante</DE></DE>

Ou encore une question sur son travail :

et qu'est ce que vous faites comme travail ?
<Turn speaker="spk1" startTime="40.394" endTime="43.041">
<DE type="pers.speaker">je suis<DE type="work.occupation"> contrôleur divisionnaire<DE type="work.occupation"> au <ENT type="org.com"> PTT </ENT></DE></DE></DE>

Ou son syndicat :

quel est votre syndicat ?
<Turn speaker="spk1" startTime="1152.961" endTime="1159.44">
<DE type="syndicat"> <ENT type="org">Force Ouvrière</ENT></DE>

Certaines indications ne découlent pas d'une question, mais se trouvent disséminées ici ou là dans la conversation. Par exemple, une question sur la connaissance du patrimoine local va introduire une indication sur la date d'arrivée à Orléans :

mais est-ce qu'il y a quelque chose dans la région ou surtout à<ENT type="loc.admi"> Orléans</ENT> que vous pouvez recommander ?
<Turn speaker="spk1" startTime="1614.656" endTime="1629.854">
oui hein ça ça dépend des goûts des personnes aussi hein vous avez des des personnes qui …
à ce moment là je peux lui expliquer je peux toujours lui dire quand même <DE type="pers.speaker"><DE type="identity.arrival"><ENT type="time.date.rel">depuis neuf ans</ENT> que je suis là</DE></DE> je commence à connaître la ville

Et la description de la recette de l'omelette peut être l'occasion de glisser son origine géographique :

comment qu'on fait une omelette ?
…
on bat tout ensemble
euh
on met dans un peu d'eau je crois
on mélange un peu d'eau
enfin on assaisonne sel poivre euh
<DE type="pers.speaker">nous en <DE type="identity.origin"><ENT type="loc.admi">Lorraine</ENT></DE></DE> on
on découpe des petits des petits morceaux de lards qu'on fait frire avant

Une évaluation de cette cascade est aussi présentée dans [Maurel et al., 2009]. La précision est de 94,2% et le rappel de 84,4%.

6 CONCLUSION

Nous avons présenté le système de cascade de transducteur CasSys qui permet une annotation des textes par un balisage. Comme CasSys utilise des graphes réalisés à l'aide de la plateforme linguistique Unitex, l'utilisateur bénéficie d'une interface graphique très conviviale pour leur réalisation.

Nous avons montré que CasSys est un système adaptable à différents types de corpus et différentes problématiques nécessitant un balisage, que ce soit de l'annotation ou de l'exploration de corpus. Signalons qu'en plus de la détection d'entités nommées et d'entités dénommantes, le système CasSys a aussi été utilisé dans le projet ANR Epac pour découper le corpus en segments syntaxiques minimaux (*chunks*) [Antoine et al., 2008].

Grâce au soutien de la Région Centre (projet Feder en cours), CasSys devrait être intégré à la plateforme Unitex d'ici la fin 2010.

7 RÉFÉRENCES

Abney S. (1996). Partial Parsing via Finite-State Cascades. *Workshop on Robust Parsing*, 8th *European Summer School in Logic, Language and Information*. Prague, Tchèquie. p. 8-15.

Antoine J. Y., Mokrane A. et Friburger N. (2008). Automatic Rich Annotation of Large Corpus of Conversational transcribed speech: the Chunking Task of the EPAC Project. Sixth language resources and evaluation conference (LREC 2008), Marrakech, Maroc, 28-30 mai.

Baude O. (2006). Corpus oraux. Guide des bonnes pratiques. Presses universitaires d'Orléans.

Courtois B. et Silberztein M. (1990). Dictionnaires électroniques du français. *Langues française*, 87. p. 11-22.

Dister A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL*. Thèse de Doctorat de Linguistique, Université catholique de Louvain.

Eshkol I. (2007). Entrer dans l'anonymat. Etude des entités dénommantes dans un corpus oral. Actes du colloque NOMINA2007. Ed. Narr, Tübingen (à paraître).

Friburger N. (2002). *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques.* Thèse de Doctorat d'Informatique, Université François Rabelais Tours.

Friburger N. et Maurel D. (2004). Finite-state transducer cascades to extract named entities in texts, *Theoretical Computer Science*, vol. 313, 94-104.

Galliano S., Gravier G. et Chaubard L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *Interspeech 2009*.

Gazeau M. A. et Maurel D. (2006). Un dictionnaire INTEX de noms de professions : quels féminins possibles ? *Cahiers de la MSH Ledoux*, p. 115-127.

Maurel D., Friburger N. et Eshkol I. (2009). Who are you, you who speak? Transducer cascades for information retrieval. *4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland, p. 220-223.

Paumier S. (2003). *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique.* Thèse de Doctorat en Informatique, Université de Marne-la-Vallée.

QUELS CORPUS POUR L'ANALYSE CONTRASTIVE ? L'EXEMPLE DES CONSTRUCTIONS VERBO-NOMINALES DE SENTIMENT EN FRANÇAIS ET EN RUSSE

Elena Melnikova[1,2], Iva Novakova[2] et Olivier Kraif[2]
1. Département des langues romanes – Université d'état d'Astrakhan (Russie)
2. LIDILEM – Université de Grenoble (Stendhal-Grenoble 3)

RÉSUMÉ

Les corpus comparables et parallèles constituent une base de données solide pour une analyse contrastive fiable. Nous nous référons aux deux types de corpus pour étudier les constructions verbo-nominales de sentiment et leurs fréquences en français et en russe. Le corpus comparable nous a servi à calculer les fréquences des noms de sentiment afin de définir la sélection des noms pour notre étude. Grâce à lui, nous avons pu également analyser les fréquences des verbes par rapport aux constructions verbo-nominales. Le corpus parallèle (ayant constitué notre corpus de contrôle) nous a permis d'établir les équivalents fonctionnels des constructions françaises en russe (par ex. les constructions impersonnelles, les verbes de sentiments, les constructions verbo-nominales de sentiment) et inversement. Ces deux types de corpus sont complémentaires en linguistique contrastive.

1 LA PROBLÉMATIQUE ET LES OBJECTIFS

L'analyse contrastive qui a pour but de décrire les similitudes et les différences entre les langues nécessite la constitution de corpus bilingues de différents types, à savoir des corpus *comparables* et des corpus *parallèles* (i.e. des corpus de traduction) (Williams 2005, Degand 2005, Celle 2006, Dalbera 2003, Lewis 2005). Cet article a pour objectif d'expliciter l'apport de ces deux types de corpus pour l'étude de la fréquence des constructions verbo-nominales de sentiment (CVN_sent) en français et en russe. Nous défendons ici l'idée que le corpus comparable et le corpus parallèle peuvent servir de base à une analyse contrastive qui vise à découvrir, à comparer et à mettre en évidence l'emploi des CVN_sent dans les deux langues.
Chaque type de corpus a ses avantages et ses inconvénients. Les corpus comparables, composés de textes originaux dans deux ou plusieurs langues qui respectent les mêmes critères de genre, de registre, de public visé, d'époque, ne permettent pas d'établir des équivalences de façon simple. Quant aux corpus de traductions (parallèles), qui offrent un accès direct aux équivalences, on leur reproche souvent de contenir des traces de la langue source et, du fait de ces artefacts traductionnels, de ne pas être totalement fiables. Ces deux types de corpus présentent donc des caractéristiques complémentaires[26] .

Selon nos premières observations et en nous référant à la recherche de G. Gak (1983, [2006]), nous avançons l'hypothèse que les CVN_sent sont moins fréquentes en russe que les verbes distributionnels de sentiment en tant qu'équivalents de la structure *Verbe+N_sent* en français. Le russe dispose d'un aspect morphologique encodé dans le verbe *(zlit'sja -> être en*

[26] Cf. aussi à ce sujet, Degand (2005).

colère et razozlit'sja -> tomber en colère) Le français, au contraire, aurait tendance à utiliser plus de CVN, du fait de l'absence d'un aspect verbal morphologique.

Nous allons vérifier notre hypothèse dans les corpus comparables et parallèles.

En premier lieu, nous étudierons le statut des deux types de corpus pour l'analyse contrastive. En deuxième lieu, nous essaierons de clarifier l'apport des deux types de corpus pour la sélection des N_sent et aussi pour l'analyse de la fréquence des équivalents fonctionnels des CVN_sent français en russe.

2 LA CONSTITUTION DES CORPUS ET LA MÉTHODOLOGIE

Pour le corpus comparable (Tableau 1) nous avons utilisé les bases de données Frantext (http://atilf.atilf.fr/frantext.htm) et Ruscorpora (http://www.ruscorpora.ru/) : 2029 textes d'auteurs contemporains entre 1950 et 2007 (environ 60 millions de mots), qui ont fourni 22 942 occurrences de CVN_sent en français et en russe.

	Français	Russe
Bases de données	Frantext	Ruscorpora
Genre	Romans, récits de voyage, essais, correspondance, mémoires	Romans, récits, essais, nouvelles, correspondance
Epoque	1950 – à nos jours	1950 – à nos jours
Nombre de mots	29 670 087	29 618 550
Nombre d'occurrences (Vsup + N_sent et N_sent+Vsup) retenues de l'échantillon	6 514 (sur 12 772)	4 853 (sur 10 170)

Tableau 1 : Corpus comparable

Pour constituer le corpus parallèle (Tableau 2) et aligner les textes, nous avons utilisé le logiciel Alinea (cf. Kraif 2006, et http://w3.u-grenoble3.fr/kraif/). Les textes parallèles (français-russes) sont des textes d'auteurs du XIX^{e} et XX^{e} siècles (libres de droits), traités dans leur version originale et leur traduction. Nous avons réuni un corpus d'environ 10 millions de mots (84 textes) qui nous a fourni 1500 occurrences de CVN_sent dans les deux langues :

	Français -Russe		Russe - Français	
Genre	Romans, récits		Romans, récits	
Epoque	1830 - 1930		1830 - 1930	
Nombre de textes	27 + 27 traductions		15 + 15 traductions	
Nombre de mots	2 007 048	2 619 016	3 674 812	2 717 174
Nombre d'occurrences (Vsup+N_sent) retenues de l'échantillon	750		750	

Tableau 2 : Corpus parallèle

L’alignement des textes russes dans *Alinea* présente certains aspects techniques spécifiques comme l’utilisation de listes de transfert (pour amorcer l’extraction des points d’ancrage) ou la translittération des textes en cyrilliques (pour identifier les noms propres et les mots apparentés).

La constitution du corpus parallèle nécessite la collecte des textes en version originale et des textes traduits et ceci dans les deux langues. Ces textes sont soumis à l’alignement (alinéa par alinéa) et sont analysés, par la suite, à l’aide de l’extraction. L'exploitation des textes parallèles (original et traduit) s’effectue en plusieurs étapes : le prétraitement avant l’alignement, l’alignement, l'élaboration des requêtes et l’analyse des résultats. Le prétraitement consiste d’abord à translittérer les textes russes. Nous utilisons ici le système le plus répandu de translittération des caractères cyrilliques russes (ISO/R 9 : 1968), repris par GOST (1983). Ce changement des cyrilliques en lettres latines nous facilite la reconnaissance des chaines apparentées (cognats ou emprunts) par le logiciel. Ensuite, si les textes téléchargés contiennent des irrégularités liées aux conversions de format (présence de macros, présence de retour-chariots à la fin de chaque ligne, etc.), il faut les éliminer en procédant à un nettoyage (qui permet de restaurer des lignes coupées, etc.). Après avoir étiqueté les textes prétraités à l’aide du logiciel Treetagger, on lance le logiciel Alinea, qui crée le projet des textes alignés. L’étape suivante est l’extraction des points d’ancrage. Ce sont les "transfuges" : des équivalents traductionnels identiques qui sont souvent nombreux comme les nombres, les noms propres, les sigles, les emprunts, les mots apparentés etc. L’étape finale est l’extraction de l’alignement complet (Figure 1).

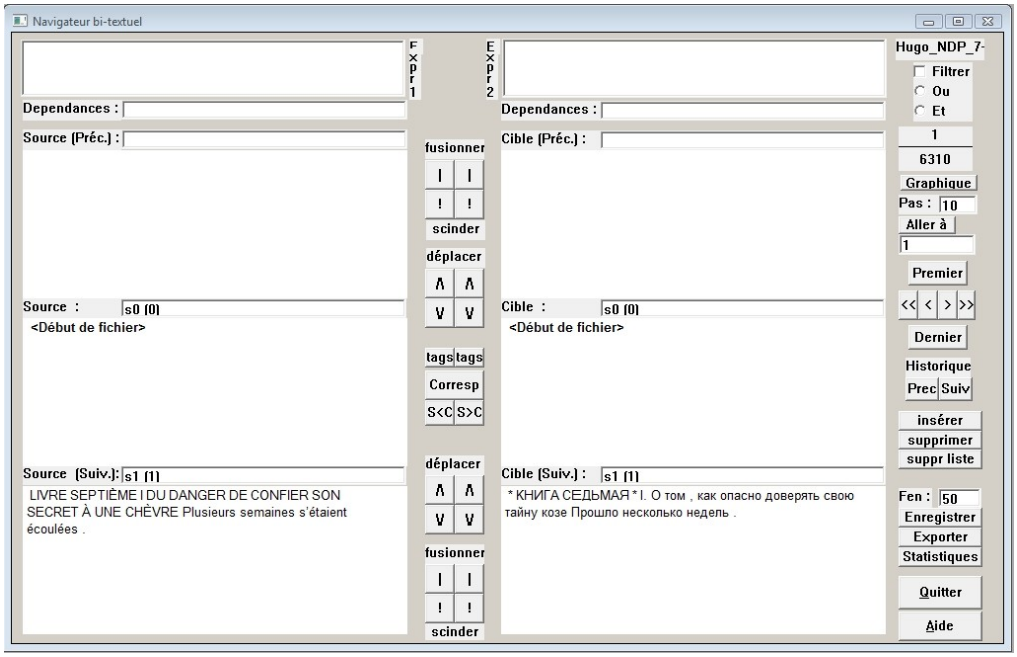


Figure 1

Le projet des textes alignés est visualisé sous forme des deux colonnes de textes français – russes, répartis en alinéas où chaque binôme (les petites fenêtres à gauche et à droite) représente une partie minimale (alinéa ou phrase) du texte original et sa traduction. Toutefois l’alignement peut donner des binômes irréguliers, ce qui demande leur ajustement (le regroupement pour mettre les segments en correspondance). Une fois l'alignement effectué, une requête peut être formulée autour d'un nom de sentiment, selon une formule intégrant des

expressions régulières, du type « </^colère.*> ». Toutes les occurrences trouvées peuvent être exportées en un seul fichier (Figure 2).

Figure 2

L'identification du patron syntaxique *V + N_sent*, qui nous a servi pour l'extraction des CVN_sent dans les textes alignés, exige une analyse minutieuse des résultats recueillis.

Le corpus de traduction, collecté grâce à Alinea, présente des avantages, ainsi que des inconvénients qui apparaissent au cours du prétraitement des textes, de l'alignement et de l'analyse des résultats. Le premier grand inconvénient est le grand nombre de manipulations de prétraitement des textes qui impose un travail à long terme. La seconde difficulté est due au fait que la majorité des textes traduits sont soumis aux droits d'auteur. La conséquence est que nous avons été obligés d'utiliser des textes du XIX^{e} et du début de XX^{e}, libres de droits. Toutefois, malgré ces difficultés, Alinea permet de mener la recherche à l'aide de différentes formules de requête, d'obtenir des résultats immédiats à travers un grand nombre de textes, de rechercher et d'extraire des constructions syntaxiquement pertinentes.

Dans la collecte des données, nous avons été aussi confrontés à d'autres problèmes comme le fait que les textes dans les deux types de corpus relèvent d'époques différentes (à cause des problèmes de droits d'auteurs) et aussi qu' ils ne sont pas de taille comparable (pour la même raison).

Conformément à notre objectif, nous allons étudier l'apport de chaque type de corpus. Nous procéderons à des analyses quantitatives des données bilingues. Dans un premier temps, nous étudierons la fréquence des N_sent sélectionnés dans les deux langues (corpus comparable). Dans un deuxième temps, nous calculerons le taux de fréquences des CVN_sent françaises par rapport aux verbes, aux CVN et aux autres constructions en russe (corpus parallèle).

3 L'APPORT DES DEUX CORPUS POUR L'ANALYSE CONTRASTIVE

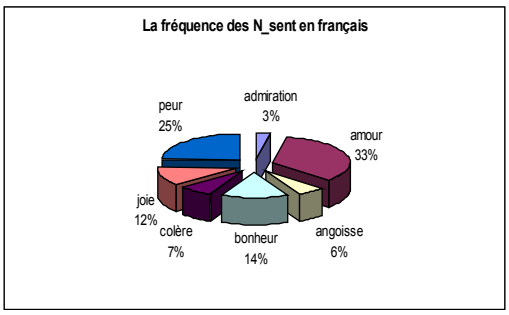
3.1 La sélection des N_sent et leurs fréquences dans les CVN (corpus comparable)

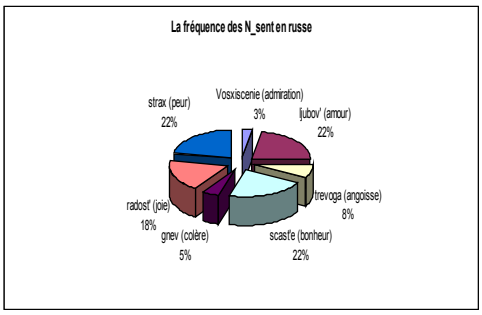
Nous avons extrait les occurrences des sept N_sent sélectionnés selon le critère de fréquence en français et en russe (*admiration / vosxiščenie, amour/ lubov', angoisse / trevoga, bonheur / sčast'ie, colère / gnev, joie / radost', peur / strax*)[27] à l'aide du patron syntaxique *V+N_sent* dans les corpus comparables monolingues Frantext et Ruscorpora (Tableau 3) :

N_sent français	**Nombre d'occ.**	**N_sent russe**	**Nombre d'occ.**
admiration	619	Vosxiščenie	767
amour	6 803	ljubov'	5 958
angoisse	1 163	trevoga	2 174
bonheur	2 815	sčast'e	5 821
colère	1 468	gnev	1 245
joie	2 324	radost'	4 929
peur	4 937	strax	6 063

Tableau 3 : La fréquence de sept N_sent (Frantext & Ruscorpora)

Le pourcentage des fréquences absolues pour *admiration* (3%) / *vosxiščenie* (3%), *angoisse* (6%) / *trevoga* (8%) et *colère* (7%) / *gnev* (5%) est presque le même dans les deux langues. Les autres noms manifestent un léger décalage dans leur fréquence, ce qui est probablement dû aux spécificités des deux corpus. Ainsi *sčast'e (bonheur)* est plus fréquent dans le russe (22%) que *bonheur* dans le corpus français (14%), tandis que *ljubov' (amour)* est moins fréquent (22%) qu'*amour* (33%) (Figures 3 et 4).

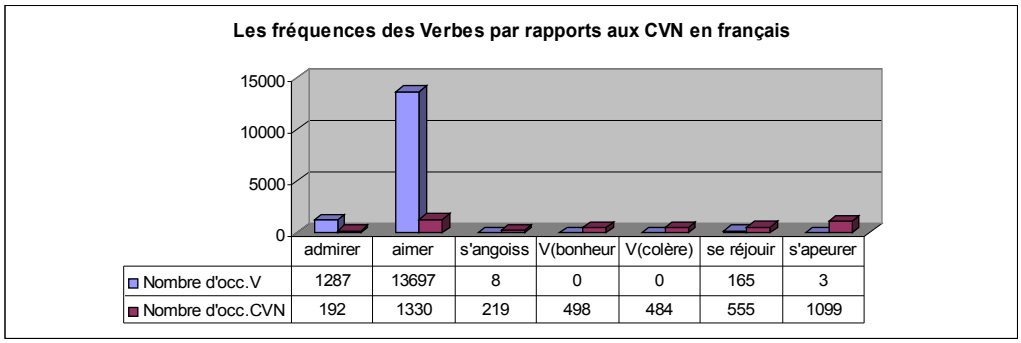




Figures 3 et 4

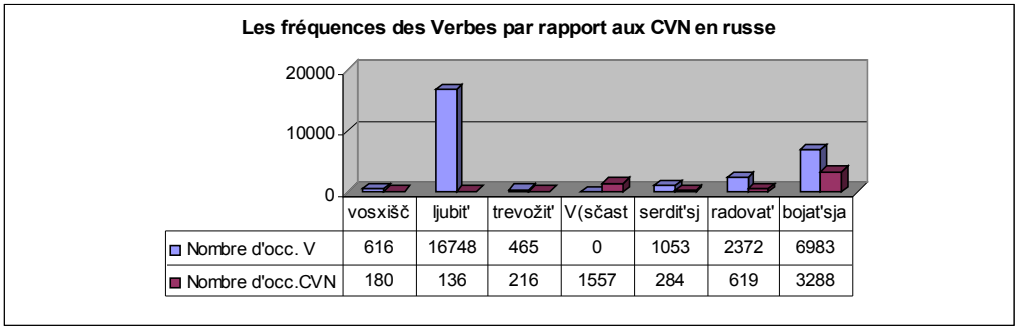
Les fréquences des CVN russes par rapport aux CVN en français dans le corpus comparable montrent que le russe utilise plus de verbes que de CVN. En français, inversement c'est les CVN qui sont plus fréquentes que les verbes pour la plupart des N_sent étudiés (Figures 5 et 6).

[27] Tous les exemples russes sont translittérés en lettres latines (ISO/R 9 : 1968, repris par GOST 1983) pour faciliter la lecture aux non-russophones.



	admirer	aimer	s'angoiss	V(bonheur	V(colère)	se réjouir	s'apeurer
Nombre d'occ.V	1287	13697	8	0	0	165	3
Nombre d'occ.CVN	192	1330	219	498	484	555	1099

Figure 5



	vosxišč	ljubit'	trevožit'	V(sčast	serdit'sj	radovat'	bojat'sja
Nombre d'occ. V	616	16748	465	0	1053	2372	6983
Nombre d'occ.CVN	180	136	216	1557	284	619	3288

Figure 6[28]

En français, sur les sept N_sent étudiés, cinq (*angoisse, bonheur, colère, joie, peur)* apparaissent le plus fréquemment dans les CVN. En revanche, en russe les verbes correspondant aux N_sent *vosxiščenie (admiration), ljubov (amour), trevoga (angoisse), gnev (colère), radost' (joie) et strax (peur),*[29] présentent des taux de fréquences plus élevées que les CVN. Ce fait confirme notre hypothèse sur l'emploi plus fréquent de verbes que de CVN en russe. Pour *sčast'e* et *bonheur* il n'existe pas de verbes correspondants dans les deux langues.

3.2 L'extraction des équivalents fonctionnels (corpus parallèle)

Nous allons maintenant tester notre hypothèse sur le corpus parallèle. Celui-ci nous a permes d'établir la liste des équivalents fonctionnels des CVN françaises en russe : CVN, verbes distributionnels et constructions impersonnelles (Tableau 4) :

[28] Pour le calcul des verbes russes on a pris en compte les verbe imperfectifs et perfectifs les plus fréquents : *vosxišat'sja / vosxitit'sja* (*admirer*), *ljubit' / poljubit'* (*aimer*), *trevožit'sja / vstrevožit'sja* (*s'angoisser*), *serdit'sja / rasserdit'sja* (« être en colère »), *radovat'sja / obradovat'sja* (*se réjouir*), *bojat'sja / ispugat'sja* (« avoir peur » qui n'a pas d'équivalents verbaux fréquents en français).

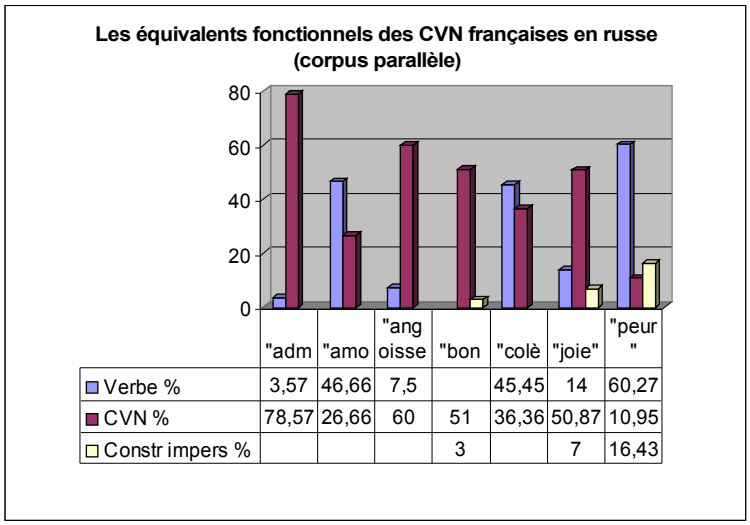
[29] On observe que *amour* et *ljubov'* sont le plus souvent utilisés dans les corpus sous forme de verbes (*aimer* et *ljubit')* et ce, dans les deux langues.

En français	**En russe**			
CVN	**Verbes distributionnels**	**CVN**	**Constructions impersonnelles**	**Autres**
Avoir de l'admiration	Vosxiščat'ja (admirer) **3,57%**	Ispytyvat' vosxiščenie (Eprouver admiration (acc)) **78,57%**	Ø	17,86%
Avoir de l'amour	Ljubit' (aimer) **46%**	Čuvstvovat' ljubov' (Sentir amour (acc)) **27%**	Ø	27%
Avoir de l'angoisse	Trevožit'sja, bojat'sja (s'angoisser) **7,5%**	Počuvstvovat' trevogu (sentir angoisse (acc)) **60%**	Ø	32,5%
Avoir le bonheur	**Ø**	Imet' sčast'e (avoir bonheur (acc)) **51%**	Emu posčastlivilos' (à lui avoir du bonheur) **3%**	46%
Être en colère	Serdit'sja (se colérer) **45%**	Byt' v gneve (être en colère (loc) **36%**	Ø	19%
Avoir de la joie	Radovat'sja (se réjouir) **14%**	Ispytyvat' radost' **51%**	Emu radostno (à lui (dat) joyeusement) **7%**	28%
Avoir de la peur	Bojat'sja (s'apeurer) **60%**	Počuvstvovat' strax (Sentir peur (acc)) **11%**	Emu strašno (à lui peureusement) **16%**	13%

Tableau 4 : Les équivalents fonctionnels des CVN françaises en russe

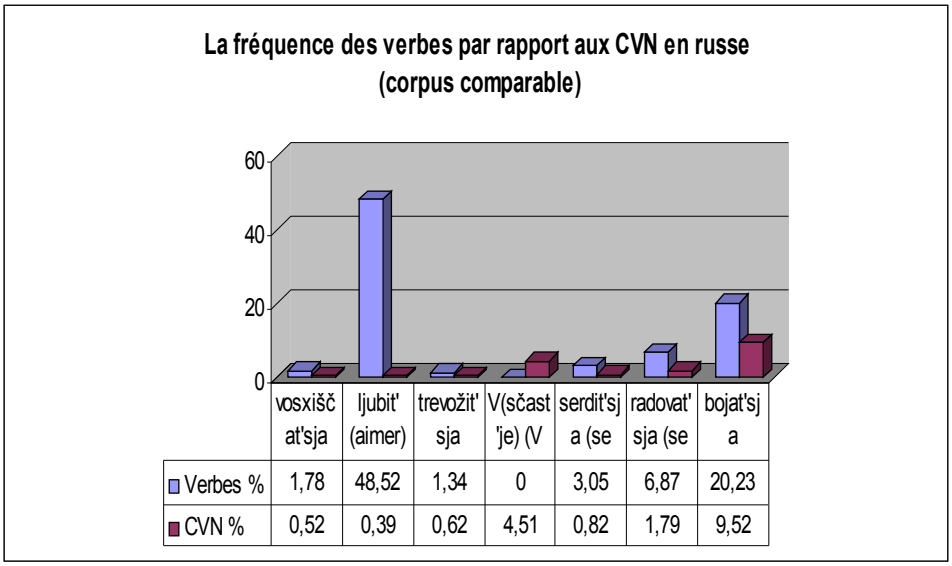
Les décomptes effectués sur le corpus de traductions (Tableau 4) montrent que *admiration* (78,57%), *angoisse* (60%), *bonheur* (51%) et *joie* (51%) apparaissent souvent dans des CVN (plus de 50%) en russe. En revanche, pour *amour, colère et peur,* le taux d'équivalents verbaux augmente sensiblement : 46% de verbes contre 27% de CVN (*amour)*, 45% de verbes contre 36% de CVN (*colère*), 60% d'équivalents verbaux contre 11% de CVN (*peur*). Le N_sent *sčast'e (bonheur),* quant à lui, fait souvent partie des CVN, car il n'existe pas d'équivalents verbaux pour ce N_sent ni en français, ni en russe. Les constructions impersonnelles ont un taux important de fréquence pour *peur* (16%) et *joie* (7%).

Si l'on compare les données des deux graphiques (figure 7 et 8), nous pouvons voir les similitudes et les divergences dans les tendances calculées à partir des différents corpus.



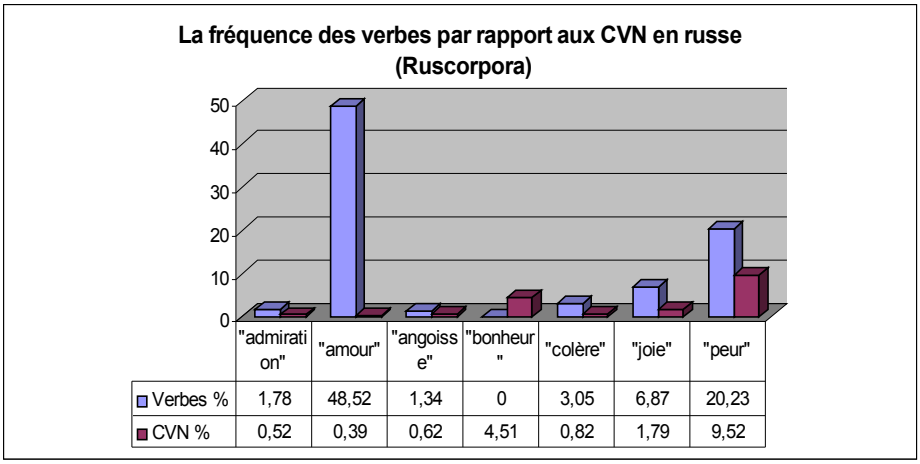
	"adm	"amo	"angoisse	"bon	"colè	"joie"	"peur"
Verbe %	3,57	46,66	7,5		45,45	14	60,27
CVN %	78,57	26,66	60	51	36,36	50,87	10,95
Constr impers %				3		7	16,43

Figure 7[30]



	vosxiščat'sja	ljubit' (aimer)	trevožit'sja	V(sčast'je) (V	serdit'sja (se	radovat'sja (se	bojat'sja
Verbes %	1,78	48,52	1,34	0	3,05	6,87	20,23
CVN %	0,52	0,39	0,62	4,51	0,82	1,79	9,52

[30] Les pourcentages des autres cas d'équivalents non-réguliers (données du Tableau 4) n'ont pas été introduits dans le graphique. Ce choix a été fait pour mieux mettre en évidence les principaux équivalents.



	"admiration"	"amour"	"angoisse"	"bonheur"	"colère"	"joie"	"peur"
Verbes %	1,78	48,52	1,34	0	3,05	6,87	20,23
CVN %	0,52	0,39	0,62	4,51	0,82	1,79	9,52

Figure 8[31]

Ainsi, le taux des fréquences élevées des équivalents verbaux pour *ljubov' (amour), gnev (colère)* et *strax (peur)* sont confirmées dans les deux corpus. En revanche, les données observées dans Ruscorpora (Figure 8) nous permettent de constater que *vosxiščenie* (*admiration*), *trevoga (angoisse)* et *radost' (joie)* ont également, en majorité, des équivalents verbaux. Ces résultats sont contrastés par rapports au corpus parallèle. Cette différence entre les données de corpus parallèle et celles du corpus russe (comparable) peut être due, à notre avis, aux aléas du corpus parallèle (les artefacts de la traduction des textes du XIX^{e} siècle, une forte influence de la langue source, les époques différentes du texte original et de sa traduction, la petite taille du corpus parallèle, ce qui augmente le risque de fluctuations liées à des phénomènes aléatoires).

Voici quelques exemples pour illustrer ce que nous venons d'exposer précédemment. La CVN *avoir peur* en français est traduite fréquemment en russe par un verbe *bojat'sja* (imperf.) ou *ispugat'sja* (perf.), (1) :

(1) Texte orig. :- *Elle eut **peur** d'être battue ;… (E. Zola. Thérèse Raquin)*
Trad. :- Она испугалась, что он побьет ее ; …
Translitt :- *Ona ispugalas', čto on pob'ët eë, ...*
Litt. :- *Elle avoir peur (passé, perf.), qu' il battre (futur, perf.) la (gén.), ...*

L'équivalent de CVN_*amour* est souvent un verbe *ljubit'* (imperf.) – *poljubit'* (perf.) ou *razljubit'* (perf. négatif) comme en (2):

(2) Texte orig. :- *C'en est fait, se dit-il avec désespoir, elle n'a plus d'amour pour moi, ... (Stendhal. La chartreuse de Parme)*
Trad. :- "Конец! - подумал он с ужасом. - Она разлюбила меня…
Translitt. :- *Konec ! – podumal on s užasom – Ona razljubila menja…*
Litt. :- *« Fin ! – penser (passé, perf.) lui (nom.) avec terreur (instr.). – Elle ne plus aimer (passé, perf.) moi (acc.)*

Le corpus parallèle nous a permis d'identifier un autre type d'équivalent russe des CVN françaises, à savoir les constructions impersonnelles. Cette construction est présentée souvent

[31] Les données de Figure 8 reflètent les données de la Figure 6. Les pourcentages de ce graphique ont été calculés automatiquement par le système Excel qui a pris pour 100% la totalité des verbes et des CVN de tous les N_sent.

par le pronom personnel au datif et un adverbe de sentiment (*mne radostno – à moi joyeusement)* comme en (3):

(3) Texte orig. : - ... *j' éprouve de la **joie** à entendre le bruit franc et net que font nos semelles sur le fond de terre dure ...(H. Barbusse. Le feu)*

Trad. :- … мне радостно слышать отчётливый звук наших шагов по твердому грунту…

Translitt. :-... *mne radostno slyšat' otčëtlivyj zvuk našix šagov po tverdomu gruntu*

Litt. :- ... *moi (dat.) joyeux entendre distinct son (acc.) nos pas (gén.) sur dur sol (dat.)...*

Pour résumer, nous avons réuni les résultats des deux corpus dans le tableau ci-dessous :

N_sent	Corpus russe (Ruscorpora)		Corpus parallèle	
	Verbe	CVN	Verbe	CVN
Vosxiščenie (admiration)	1,78% +	0,52% -	3,57% -	78,57% +
Ljubov' (amour)	48,52% +	0,39% -	46,66% +	26,66% -
Trevoga (angoisse)	1,34% +	0,62% -	7,5% -	60% +
Sčast'e (bonheur)	0	4,51%	0	51%
Gnev (colère)	3,05% +	0,82% -	45,45% +	36,36% -
Radost' (joie)	6,87% +	1,79% -	14% -	50,87% +
Strax (peur)	20,23% +	9,52% -	60,27% +	10,95 -

Tableau 5 : Taux des verbes et des CVN dans les deux corpus en russe

Dans le corpus parallèle trois N_sent (*ljubov (amour), gnev (colère) et strax (peur))* apparaissent plus fréquemment sous la forme du verbe correspondant, plutôt qu'au sein d'une CVN. Les trois autres *vosxiscenie (admiration), trevoga (angoisse) et radost' (joie)* présentent des fréquences élevées dans des CVN par rapport aux verbes correspondants[32]. Nous expliquons ce résultat contrasté surtout par l'influence sur la traduction russe du texte source français, où les N_sent apparaissent souvent sous forme de CVN. Le corpus comparable russe révèle, comme nous l'avons vu, une prédominance de verbes et ce, pour pratiquement tous les N_sent étudiés.

4 CONCLUSIONS

Les corpus comparables et parallèles sont complémentaires et présentent plusieurs avantages, mais aussi des inconvénients (Tableau 6) :

[32] Nous avons enlevé ici *bonheur* de nos analyses car il n'existe pas de verbes correspondants dans les deux langues pour ce nom.

Corpus comparables		Corpus parallèles	
+	-	+	-
L'accès de manière directe aux occurrences de N_sent, la constitution des corpus de taille importante et le **calcul des fréquences** d'un N_sent, des verbes et des CVN.	ne permettent pas d'établir des équivalents fonctionnels	permettent **d'établir les équivalents fonctionnels** des CVN_sent français en russe et inversement	La langue des textes traduits contient souvent des traces de la langue source

Tableau 6 : La complémentarité des deux corpus

L'hypothèse émise est vérifiée en partie. Selon le corpus comparable, six N_sent (à l'exception *bonheur* et *sčast'e*) ont des équivalents verbaux en russe. En revanche, selon le corpus parallèle, c'est uniquement trois N_sent (*ljubov' (amour), gnev (colère)* et *strax (peur)*) ont des équivalents verbaux en russe par rapport aux CVN françaises. Le russe, malgré l'aspect verbal morphologisé, utilise aussi bien des CVN_sent pour *vosxiščenie (admiration), trevoga (angoisse), bonheur (sčast'e), radost' (joie)*. Ceci est surtout révélé par le corpus parallèle, ce qui peut être dû à l'influence de la langue source. Par ailleurs, il s'ensuit que le russe, malgré l'aspect verbal morphologisé, emploie aussi bien les verbes de sentiment que les CVN_sent. Les spécificités des deux langues se révèlent à travers les calculs sur les deux corpus: 1) le corpus comparable et le corpus parallèle ont laissé découvrir les équivalents « absolus » des trois N_sent (*ljubov' (amour), gnev (colère), strax (peur)*) ; 2) le corpus parallèle a permis d'identifier un autre type d'équivalents, les constructions impersonnelles en russe.

Les résultats provenant des deux types de corpus ne sont pas identiques et peuvent être liés à de nombreuses causes : 1) l'artefact de la traduction ; 2) une forte influence de la langue source ; 3) les époques différentes des textes de deux corpus ; 4) les époques différentes des textes originaux et de leurs traductions dans le corpus parallèle ; 5) la petite taille du corpus parallèle, plus difficile à obtenir du fait des droits d'auteur liés à la traduction ; 6) les textes issus d'un seul genre littéraire, et un large emploi de temps de narration (l'imparfait et le passé simple) ; 7) les systèmes différents du français et du russe.

La comparaison des résultats des deux types de corpus nous a amenés à réfléchir sur le statut des équivalents fonctionnels établis: sont-ils des épiphénomènes liés à l'idiosyncrasie de l'auteur/traducteur (« posture sourcière ») ou bien des correspondances relevant du système des deux langues sur le plan sémantique, syntaxique et morphologique. Nous estimons, à l'état actuelle de notre recherche que certes, l'influence de la « posture sourcière » est assez forte, mais que certainement l'hypothèse avancée suite aux recherches de Gak (1983, [2006]) et à nos propres intuitions mérite d'être nuancée.

Cette étude laisse encore des pistes à explorer. Notamment, vérifier la fiabilité des données issues du corpus parallèle en croisant les données dans les deux directions (du français vers le russe et du russe vers le français), et observer les fréquences d'emploi de différents équivalents fonctionnels dans les corpus comparables, par un va-et-vient constant entre les deux types de corpus. En outre, il serait également intéressent d'observer l'apport des deux corpus pour l'étude de l'aspectualité des N_sent au sein des CVN_sent. Le calcul des paramètres aspectuels (l'aspect lexical et grammatical des verbes, et aussi les classifieurs, les modifieurs et les déterminants des N_sent) sera effectué à partir de la combinatoire syntaxique et lexicale des CVN_sent dans les deux langues.

5 RÉFÉRENCES

Celle A. (2006). *Temps et modalité. L'anglais, l'allemand, le français en contraste*. Allemagne : Presses scientifiques européennes.

Dalbera J-Ph**.** « Le corpus entre données, analyse et théorie », *Corpus* [En ligne], n°1 | novembre 2002, mis en ligne le 15 décembre 2003, Consulté le 24 juin 2009. URL : http://corpus.revues.org/index10.html

Degand L. (2005). « De l'analyse contrastive à la traduction : Le cas de paire puisque – aangezien ». *La linguistique de corpus.* Rennes : Presses universitaires de Rennes. p. 155-168.

Gak V. ([1983] (2006). *Sravnitelnja tipologija francuzskogo i russkogo jazykov (La typologie contrastive du français et du russe).* Moskva : URSS.

Kraif O. (2006). « Qu'attendre de l'alignement de corpus multilingues ? ». *Revue Traduire, 4e Journée de la traduction professionnelle*, Société Française des Traducteur, N° 210, p. 17-37.

Lewis D. M. (2005). « Corpus comparable et analyse contrastive : l'apport d'un corpus de français / anglais de discours politique à l'analyse des connecteurs adversatifs ». Dans G. Williams *La linguistique de corpus.* Rennes : Presses universitaires de Rennes.

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

Frantext. CNRS, ATILF (Analyse et traitement informatique de la langue française), UMR CNRS-Université Nancy2, http://atilf.atilf.fr/categ.htm

Ruscorpora. (www.ruscorpora.ru)

DES QUESTIONS LINGUISTIQUES SOULEVÉES PAR LES RÉSULTATS D'ALIGNEMENT DES MOTS *KATAKANA*

Yayoi Nakamura-Delloye
Laboratoire LeSCLaP – Université de Picardie

RÉSUMÉ

Cette communication présente les réflexions que nous avons eues lors de l'évaluation des résultats de l'alignement des mots *katakana*. Dans le cadre de nos travaux antérieurs sur le développement d'un système d'alignement des phrases (Nakamura-Delloye 2005), nous avons conçu une méthode d'alignement de ces mots en *katakana* basée sur leur retranscription en alphabet latin. Afin d'approfondir nos études, nous avons réalisé une évaluation plus complète de cette technique. Mais lors de l'évaluation, nous avons été confrontés à des questions linguistiques remettant en cause l'évaluation même des résultats. En effet, bien que dans certains cas la justesse ou la fausseté de la segmentation soit évidente, dans d'autres cas nous n'arrivons pas à trancher.

1 INTRODUCTION

Le présent article présente les réflexions que nous avons eues lors de l'évaluation des résultats de l'alignement des mots *katakana*. Le syllabaire *katakana* est l'un des trois principaux systèmes d'écriture utilisés dans les textes japonais, qui est généralement employé pour les mots emprunts qu'il sert à transcrire phonétiquement. Dans le cadre de nos travaux antérieurs sur le développement d'un système d'alignement des phrases (Nakamura-Delloye 2005), nous avons conçu une méthode d'alignement de ces mots en *katakana* basée sur leur retranscription en alphabet latin. L'efficacité de ce type de méthode était déjà connue (Collier *et al*. 1997) (Brill *et al*. 2001) (Tsuji 2001) (Tsuji *et al.* 2002) mais nos travaux différaient des travaux connexes notamment par la non-utilisation d'analyseur morphologique et par la retranscription par transducteur. Afin d'approfondir nos études, nous avons réalisé une évaluation plus complète de cette technique, et nous avons alors été confrontés à des questions linguistiques remettant en cause l'évaluation même des résultats.

L'article décrit d'abord une méthode de segmentation des mots composés en *katakana* (§ 2) ainsi qu'une méthode de mise en correspondance de ces mots avec leur traduction (§ 3). Nous nous intéresserons ensuite au résultat d'expérience — corpus, évaluation et analyse —, et surtout aux questions rencontrées lors de cette évaluation (§ 4).

2 SEGMENTATION DES MOTS COMPOSÉS EN *KATAKANA* À L'AIDE D'UN ARBRE LEXICOGRAPHIQUE

Comme le japonais ne possède pas de signes permettant de segmenter les phrases *a priori*, un système d'analyse morphologique est généralement utilisé, système dont l'objectif est de transformer, à l'aide d'un dictionnaire et de probabilité des combinaisons, la phrase en une suite d'unités appartenant chacune à une catégorie morpho-syntaxique.

2.1 Segmentation par analyseur morphologique

L'analyse morphologique commence par la consultation d'un dictionnaire afin de trouver toutes les séquences qui peuvent constituer une unité morphologique. Par exemple, pour une séquence « *zen-koku-to-dô-fu-ken-gi-chô-kai* » (cf. Fig. 1), on peut trouver 18 candidats d'unités morphologiques composantes en consultant un dictionnaire. Ces candidats donnent ensuite lieu à différentes combinaisons possibles présentées dans la figure 1. L'opération consiste ensuite à déterminer la combinaison la plus adéquate de ces candidats, généralement avec une méthode probabiliste, pour

trouver la segmentation de la séquence. Mais cette méthode a comme inconvénient d'être dépendante de l'existence et de la qualité des dictionnaires.

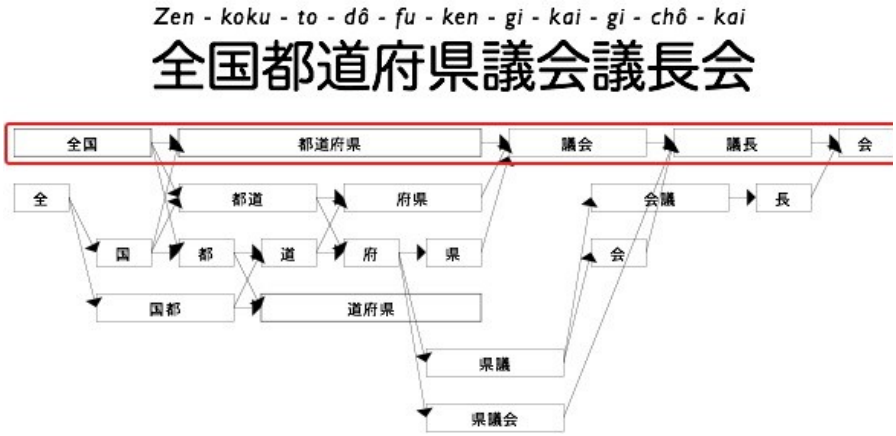


Figure 1: Possibilités de combinaison de tous les candidats

Toutefois, il existe également une méthode classique d'analyse morphologique partielle permettant d'extraire les éléments lexicaux sans aucune connaissance extérieure, appelée segmentation par type de caractère, qui met à profit la principale particularité de l'écriture japonaise utilisant plus de trois systèmes d'écriture.

2.2 Segmentation par type de caractère et ses problèmes

En japonais, plusieurs types de caractères sont utilisés selon la nature des mots. Il existe trois principaux systèmes d'écriture : *kanji*, *hiragana* et *katakana* (cf. Fig. 2). Les *kanji* sont des idéogrammes qui sont utilisés pour représenter les mots pleins et les radicaux qui ont un sens. Les *hiragana* sont un des deux syllabaires japonais et ils sont souvent utilisés pour représenter les mots grammaticaux (notamment les particules) et la partie variable des mots variables (verbes, adjectifs). Les *katakana* sont l'autre syllabaire japonais et comme je l'ai déjà évoqué dans l'introduction, ils sont généralement employés pour transcrire phonétiquement les mots emprunts.

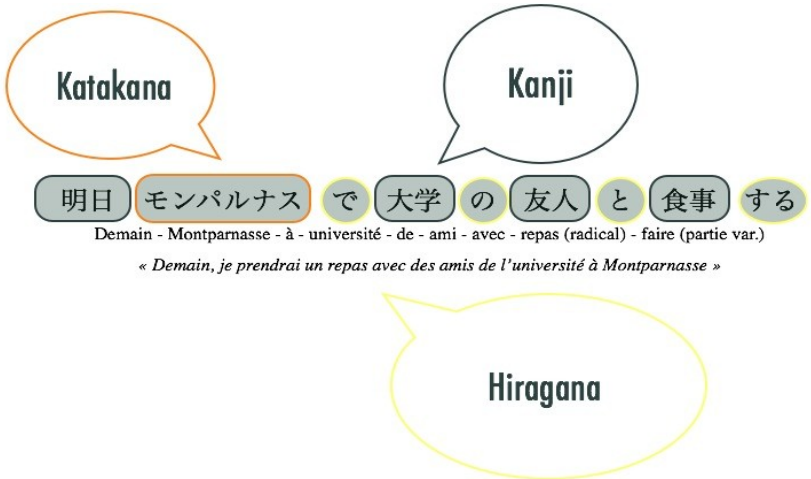


Figure 2: Les trois systèmes d'écriture japonais

Cette méthode de segmentation possède toutefois deux grands problèmes : impossibilité de segmentation des séquences de mots composés constituées de plusieurs substantifs du même type de caractère, juxtaposés les uns derrière les autres (cf. Fig. 3 : deuxième exemple « *rin-ji-koku-kai-hei-kai* » et troisième exemple « *yu-u-za-i-n-ta-a-fu-e-i-su* ») ; segmentation fausse des mots constitués de différents types de caractères (cf. Fig. 3 : premier exemple « *i-ki* »).

Dans le cadre du développement d'un système d'alignement des phrases traitant les textes japonais sans analyseur ni dictionnaire, nous avons conçu une amélioration de la segmentation par type de caractère en résolvant le premier type de problème, à savoir la segmentation des séquences où la frontière entre les deux mots n'est pas marquée par un changement de type de caractère.

2.3 Amélioration de la segmentation par type de caractère

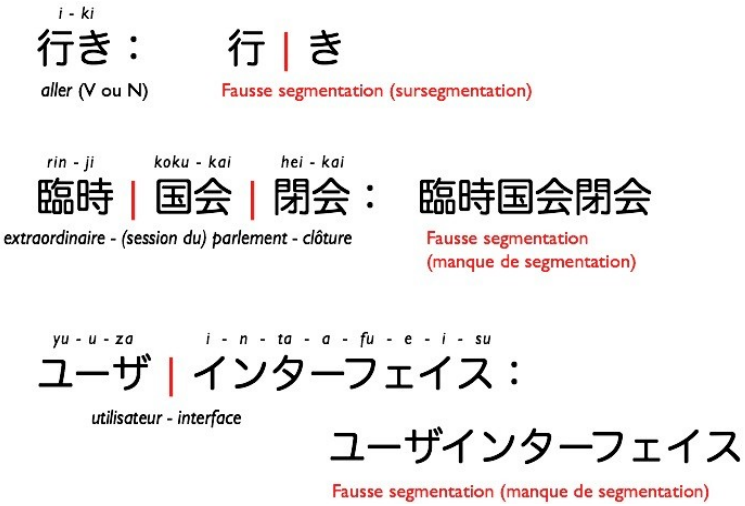


Figure 3 : Problèmes de la méthode de segmentation par types de caractère

Cette amélioration est inspirée par la méthode d'analyse morphologique partielle proposée par (Kay & Röscheisen, 1993). Elle consiste à trouver les sous-chaînes préfixales ou suffixales communes à plusieurs formes effectives des mots graphiques et à déterminer les radicaux, porteurs de sens. Il s'agit donc de la recherche des sous-chaînes préfixales communes à plusieurs formes effectives. Par exemple, à partir de trois séquences « engagé », « engager » et « engageons », nous pouvons obtenir la sous-chaîne préfixale commune « engag ».

Nous avons adapté cette méthode au traitement des mots composés en *kanji* et en *katakana*. En effet, trouver les frontières de mots dans une séquence constituée d'un même type de caractère est également une recherche des sous-chaînes communes à plusieurs formes effectives. Par exemple, à partir de trois séquences « *shoku-ryô-kyô-kyû* », « *shoku-ryô-bu-soku* », « *shoku-ryô-ki-ki* » de l'exemple présenté dans la figure 4, nous pouvons obtenir la sous-chaîne préfixale commune « *shoku-ryô* ». L'adaptation de cette approche au traitement des textes japonais a nécessité quelques modifications, notamment le traitement des parties restantes considérées comme des suffixes, qui doivent, dans le cas du japonais, être conservées en tant qu'unités lexicales autonomes. On obtient donc à partir d'un mot graphique *a b*, non pas uniquement une forme de base *a*, mais deux formes de base *a* et *b*. Dans le cas de l'exemple cité précédemment, on obtient à partir des trois séquences, non pas un lemme, mais quatre lemmes « *shoku-ryô* », « *kyô-kyû* », « *bu-soku* » et « *ki-ki* ».

Nous avons également appliqué la même méthode à la segmentation des séquences en *katakana*. Mais, dans le cas des séquences de *katakana*, nous ne cherchons pas de sous-chaînes communes à plusieurs séquences de *katakana*, mais des sous-chaînes semblables à une autre séquence ou à un lemme extrait d'une séquence plus grande. Autrement dit, lorsqu'on a une séquence « *ba-i-po-o-ra-a-to-ra-n-ji-su-ta-a* », c'est uniquement quand on trouve « *to-ra-n-ji-su-ta-a* » utilisé seul qu'elle est segmentée en deux mots (cf. Fig. 5, exemple du haut). Ainsi, on empêche la segmentation des séquences telles que « *i-n-su-to-o-ru* » et « *i-n-to-ro-da-ku-shi-yo-n* » en deux parties, même si la

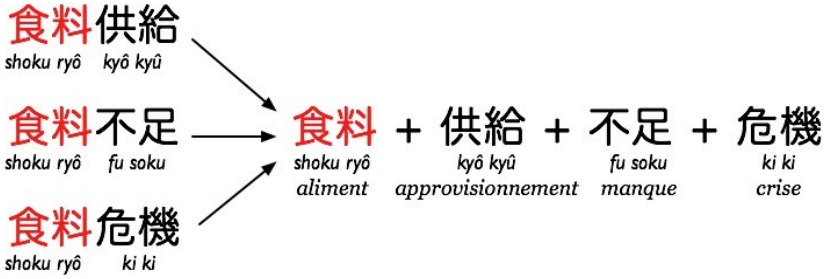


Figure 4 : Segmentation par recherche des sous-chaînes communes

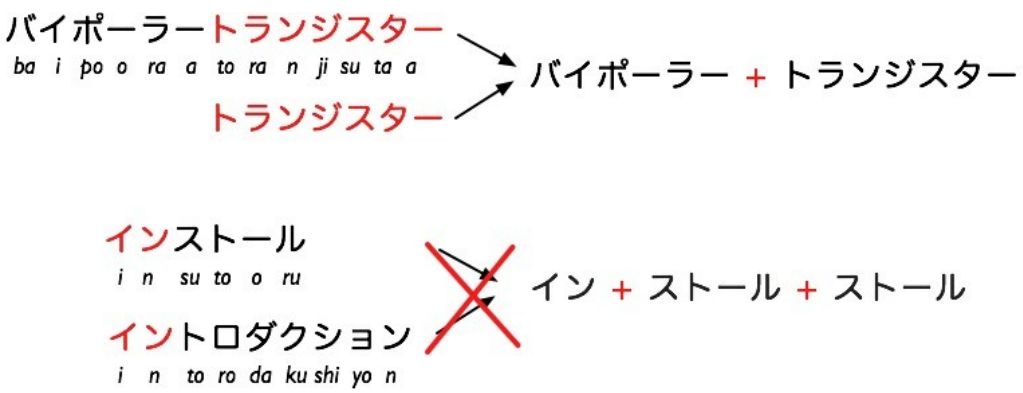


Figure 5 : Segmentation de séquences de katakana

séquence « *i-n* » est la sous-chaîne commune de ces deux formes (cf. Fig. 5, exemple de bas).

Cette méthode peut être implémentée efficacement à l'aide de la structure de données appelée arbre lexicographique. La figure 6 représente un exemple d'arbres qui vérifient des chaînes préfixales et suffixales. L'arbre gauche est construit avec des séquences normales et il est destiné à trouver les sous-chaînes préfixales. L'arbre droit est construit avec des séquences inversées et il est destiné à trouver les sous-chaînes suffixales.

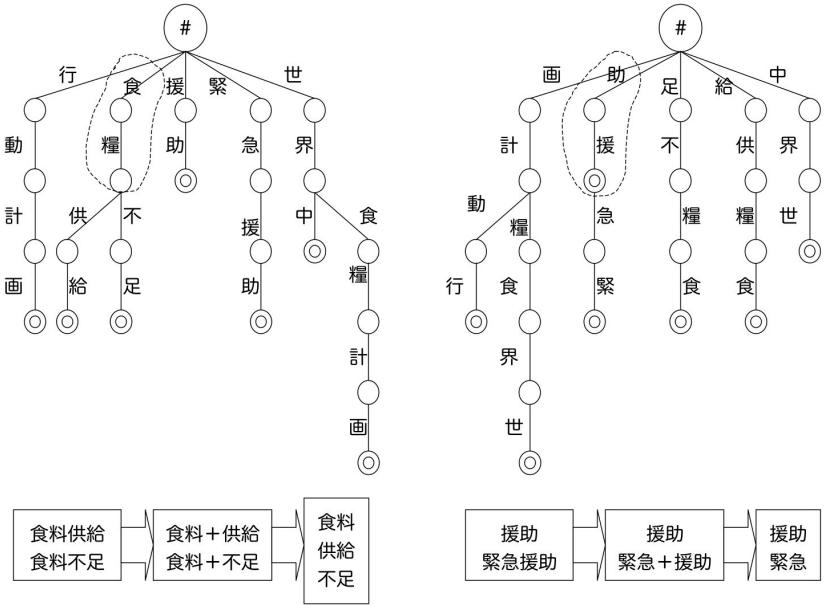


Figure 6 : Recherche des sous-chaînes communes avec des arbres lexicographiques

3 ALIGNEMENT DES MOTS *KATAKANA* PAR LA RETRANSCRIPTION AVEC TRANSDUCTEUR

Nous nous intéressons à présent à la méthode dans laquelle les mots katakana ainsi extraits sont alignés avec leur mot d'origine à partir d'un corpus parallèle non aligné. Leur alignement consiste en leur retranscription en alphabet latin à l'aide d'un transducteur spécifiquement conçu.

La figure 7 représente la procédure d'appariement d'un mot en *katakana*. Les mots en *katakana* extraits avec la méthode de segmentation constituent la liste des mots *katakana* (cf. étape 1 Fig. 7). Notre transducteur retranscrit chaque mot en *katakana* en plusieurs formes en alphabet latin (cf. étape 2 Fig. 7). Avec l'autre corpus, on constitue également une liste des mots (cf. étape 3 Fig. 7).

Chaque mot de cette liste est comparé avec des retranscriptions des mots *katakana* (cf. étape 4 Fig. 7) et si la similarité entre le mot considéré et une séquence de retranscription d'un mot en *katakana* atteint un seuil prédéfini, ce mot est stocké dans la liste des candidats (cf. étape 5 Fig. 7). Une fois l'examen des mots terminé, on cherche, parmi les candidats extraits, le mot d'origine le plus probable pour chaque mot japonais en *katakana*, et on constitue la liste des mots en *katakana*

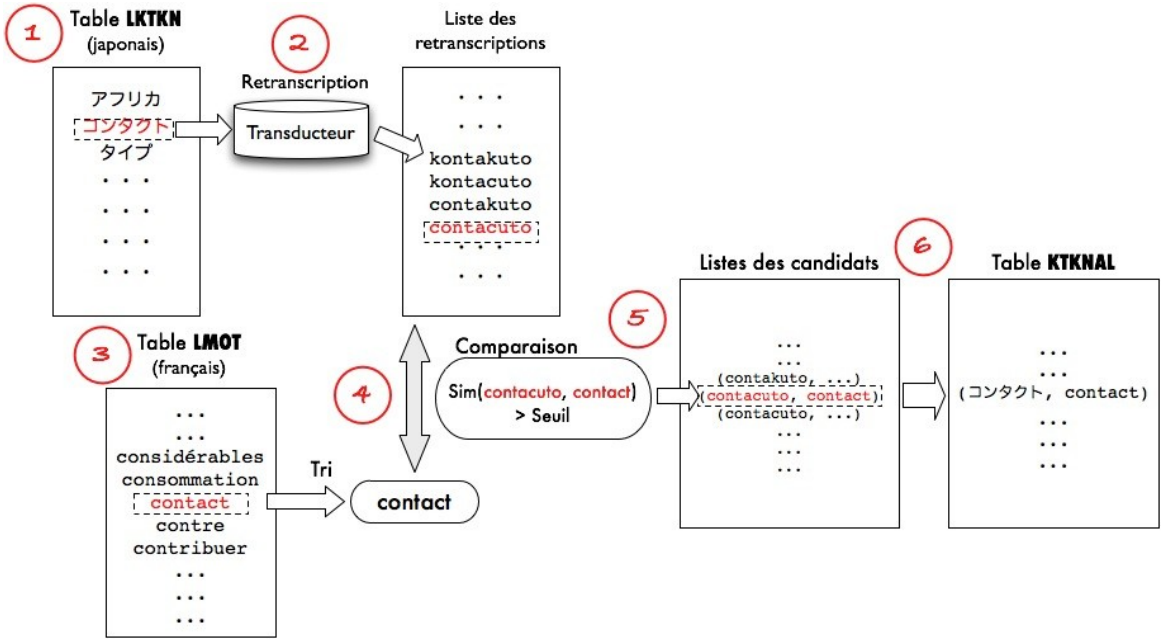


Figure 7 : Procédure d'alignement des mots katakana

alignés (cf. étape 6 Fig. 7).

Afin de permettre la retranscription, une grammaire a d'abord été définie. Cette grammaire a ensuite été transformée en un ensemble de règles de transition et de sortie pour constituer un transducteur de retranscription. Les règles de transition indiquent la transition d'un état à un autre provoquée par chaque symbole d'entrée. Les règles de sortie indiquent un ou plusieurs symboles de sortie liés à chaque état. À l'aide de ce transducteur, un mot *katakana* est retranscrit en plusieurs formes possibles.

Ces formes retranscrites sont ensuite comparées avec tous les mots lexicaux du texte de l'autre langue pour calculer leur similarité de forme. La similarité doit être calculée tout en tenant compte d'éventuelles divergences de forme dues à la différence de système phonétique / phonologique entre le japonais et l'autre langue. À cet effet, nous recourons aux méthodes de calcul utilisées pour la mise en correspondance des cognats largement étudiées dans le cadre de travaux sur l'alignement entre les textes de langues apparentées. Notre formule est inspirée notamment de celle de la sous-chaîne maximale parallèle proposée par Kraif (2001). Toutefois, du fait des besoins particuliers de la retranscription des *katakana*, elle diffère de cette dernière par le fait qu'elle tient compte non seulement de la sous-chaîne maximale mais aussi des consonnes communes. Le nombre de consonnes communes est pris en compte pour favoriser les deux chaînes ayant le plus de caractères consonantiques communs plutôt que celles dont les caractères vocaliques coïncident le plus.

À cette similarité, nous combinons également la similarité de distribution afin d'exclure les correspondances hasardeuses et déterminons pour chaque mot *katakana* le mot de l'autre texte ayant la similarité la plus élevée et supposé en relation de traduction.

4 EXPÉRIENCE : CORPUS, ÉVALUATION, ANALYSE

Afin d'approfondir nos études antérieures, nous avons réalisé à nouveau une évaluation plus

complète de ces techniques de segmentation et de mise en correspondance des mots *katakana*.

4.1 Corpus

Nous avons utilisé des manuels (anglais/français/japonais) de produits électroniques de la société Apple, disponibles sur le site de la société. Pour l'alignement français-japonais, nous avons utilisé en plus un corpus constitué d'articles du Monde Diplomatique (LMD ci-dessous). Ces corpus sont caractérisés par les occurrences importantes des *katakana*. À titre d'exemple (cf. Tab. 1), l'extrait du roman japonais « *La fin des temps* » de Haruki Murakami (Shinchosha, 1985) et une série d'articles du quotidien japonais « *Yomiuri* » contenaient respectivement 7 et 4% de mots *katakana* dans l'ensemble des mots utilisés, tandis que les articles traduits du Monde Diplomatique en comportaient 12%, et les manuels Apple, eux, 17%.

Le corpus Apple — comme tous les textes de ce domaine — est caractérisé par l'importance des mots emprunts, souvent des néologismes, appartenant notamment au vocabulaire informatique. En revanche, le corpus LMD contient plus de noms propres transcrits par ce syllabaire.

	Roman	Journal	LMD	Apple
Nb de mots	20 150	22 367	24 343	63 638
Nb de mots différents (A)	2 857	3 052	4 290	2 272
Nb de mots katakana (B)	201	120	529	384
B / A	**7 %**	**4 %**	**12 %**	**17 %**

Table 1 : Caractéristiques des corpus

4.2 Résultat : segmentation des séquences en katakana

Les corpus ont d'abord été segmentés en mots par deux méthodes différentes : avec notre méthode de segmentation et par un analyseur morphologique du japonais ChaSen, analyseur morphologique utilisé très largement au Japon et développé par le Nara Institute of Science and Technology.

ChaSen a tendance à segmenter plus, fournissant 10% de mots en plus. En effet, ChaSen découpe très souvent les mots absents dans son dictionnaire en sous-chaînes qui y figurent. Ainsi beaucoup de noms propres et de néologismes ont été sursegmentés de manière erronée.

Dans le résultat de notre méthode de segmentation, il y a beaucoup moins de sursegmentations erronées. En revanche, on constate plus l'absence de segmentation. Par exemple, ChaSen a segmenté les séquences telles que « *su-te-e-ta-su-ra-m-pu* » (status light) et « *ka-a-ki-t-to* » (car kit) en deux mots alors que notre méthode n'a effectué aucun découpage.

Il y a également des cas inverses. Par exemple, les séquences telles que « *pu-re-i-ri-su-to* » (play list) et « *de-e-ta-ro-o-mi-n-gu* » (data roaming) ont été segmentées en deux mots par notre méthode, mais pas par l'analyseur ChaSen.

Néanmoins, notre plus grand problème résidait dans l'évaluation de ce résultat de segmentation. En effet, bien que dans certains cas la justesse ou la fausseté de la segmentation soit évidente, dans d'autres cas nous n'arrivons pas à trancher : « *su-ra-i-do-sho-o* » (slideshow) n'est-il qu'un seul mot ? Et « *sa-i-n-a-p-pu* » (sign up) ou « *ba-k-ku-a-p-pu* » (backup/back up) ? Le figement d'un mot est difficile à déterminer d'autant plus sans doute lorsqu'il s'agit d'un mot emprunt d'une autre langue.

Ainsi nous sommes retombés dans la question classique de la définition de mot. Nous avons donc décidé de ne pas évaluer à cette étape et de regarder l'influence de ces segmentations différentes dans le résultat de l'alignement.

Mots d'origine	Segmentation par analyseur	Segmentation par recherche des sous-chaînes
Car kit	☺ カー | キット	カーキット
Broadband	ブロード | バンド	☺ ブロードバンド
Data roaming	データローミング	☺ データ | ローミング
Playlist	☺ プレイリスト	プレイ | リスト

☺ = segmentation ayant permis l'alignement des mots

Table 3 : Exemples de résultat de segmentation des séquences en katakana

4.3 Résultat : appariement des mots katakana

Le nombre de paires correctement alignées est *grosso modo* semblable pour les deux méthodes (cf. Tab. 2). En effet, la segmentation (ou la non segmentation) peut à la fois favoriser et défavoriser l'alignement. Dans certains cas, la segmentation en plus petites unités favorise l'alignement, et dans d'autres elle l'empêche. Par exemple (cf. Tab. 3), « *ka-a-ki-t-to* » (car kit) et « *bu-ro-o-do-ba-n-do* » (broadband) sont segmentés en deux mots par l'analyseur, mais la segmentation a empêché l'alignement du deuxième exemple contrairement au premier qui a été aligné avec les deux mots anglais correspondants. « *pu-re-i-ri-su-to* » (playlist) et « *de-e-ta-ro-o-mi-n-gu* » (data roaming) sont segmentés en deux mots par notre méthode et non par l'analyseur. Là encore, la segmentation de « *pu-re-i-ri-su-to* » a empêché l'alignement alors que la segmentation en « *de-e-ta* » et « *ro-o-mi-n-gu* » a permis l'alignement avec le mot anglais correspondant.

La conséquence intéressante de ce résultat disparate est que malgré leur résultat comparable, chaque méthode contient plus de 10% de paires qui ne sont pas alignées par l'autre méthode.

Pour l'alignement également, nous avons comparé deux méthodes, l'une avec la similarité calculée par la similarité des formes et des distributions et l'autre sans prise en compte des distributions. Dans des travaux antérieurs de mise en correspondance des mots *katakana*, la distribution n'est pas prise en compte, mais elle améliore sensiblement la précision. Dans le cas de l'alignement anglais-japonais, nous avons obtenu une précision de 76-82% avec un taux de rappel de 50-55%. Avec la prise en compte des distributions, ce taux de précision remonte à 96-98%, mais le rappel tombe à moins de 25%. Le résultat d'alignement français-japonais du corpus Apple est décevant avec une précision de 45% et un taux de rappel de 25%. Mais avec la prise en compte des distributions, la précision remonte jusqu'à 88% avec un rappel de 16%. Le résultat d'alignement du corpus LMD est tout de même encourageant avec une précision de 93% et un taux de rappel de 34%.

5 CONCLUSION ET PERSPECTIVES

Le résultat de segmentation n'a pas montré l'avantage d'une des deux méthodes comparée à l'autre. La segmentation (ou la non-segmentation) peut à la fois favoriser et défavoriser l'alignement. Cette expérience nous a amené surtout à remettre en cause la définition même de mot, que nous avons choisi comme unité d'alignement. Nous avons donc mis de côté l'évaluation à cette étape pour regarder plutôt l'influence de ces segmentations sur le résultat de l'alignement.

Le résultat d'alignement n'a pas non plus montré l'avantage d'une de ces deux méthodes de

segmentation. La segmentation (ou la non segmentation) peut à la fois favoriser et défavoriser l'alignement. La solution serait peut-être de combiner les deux méthodes de segmentation pour profiter entièrement des paires de mots alignés par au moins une des deux méthodes.

Quant à l'opération d'alignement, notre résultat n'est pas encore totalement satisfaisant. En effet, notre grammaire de retranscription définie manuellement est incomplète. En revanche, les travaux existants (Tsuji 2001) (Tsuji *et al.* 2002) proposent une méthode pour définir automatiquement une grammaire à partir de ressources terminologiques bilingues. Sur le modèle de ces travaux et par la prise en compte de la probabilité de chaque transcription, nous pourrions certainement améliorer nos résultats. Par ailleurs, l'exploitation de corpus comparables, présentant différents avantages par rapport aux corpus parallèles — notamment leur importance et leur diversité —, a attiré depuis toujours les chercheurs travaillant dans le domaine de la constitution automatique de ressources multilingues et plusieurs travaux portant sur l'extraction de terminologies bilingues ont été présentés, et ce non seulement avec des approches pour les langues parentes (Dejean et Gaussier 2002) (Morin et Daille 2004), mais aussi celles adaptées pour le japonais (Kageura *et al.* 2000). L'intérêt de l'utilisation de notre méthode d'alignement des mots *katakana* dans ce type de travaux exploitant des corpus comparables est également incontestable.

6 RÉFÉRENCES

Brill E., Kacmarcik G. et Brockett C. (2001). « Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs ». Asia Federation of Natural Language Processing. p. 393-399.

Collier N., Hirakawa H. et Kumano A. (1997). « Acquisition of English-Japanese proper nouns from noisy-parallel newswire articles using KATAKANA matching ». Proceedings of the Natural Language Processing Pacific Rim Symposium 1997. p. 309-314.

Dejean H. et Gaussier E. (2002) « Une nouvelle approche à l'extraction de lexique bilingues à partir de corpus comparables ». Dans W. Teubert et R. Krishnamurthy (éd.), Corpus Linguistics : Critical Concepts in Linguistics. Routledge Publishers.

Morin E. et Daille B. (2004). « Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé », Vol. 45/3. p.103-122

Kageura K., Tsuji K. et Aizawa N. (2000). « Automatic Thesaurus Generation through Multiple Filtering ». Proceedings of the 18th Conference on Computational linguistics, Vol. 1. p. 397-403.

Kay M. et Röscheisen M. (1993). « Text-translation alignment ». Computational Linguistics, 19 (1), p. 121-142.

Kraif O. (2001). « Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation ». TAL, 42 (3).

Nakamura-Delloye Y. (2005). « Système AlALeR : Alignement au niveau phrastique des textes parallèles français-japonais ». Actes de la conférence TALN/RECITAL 2005. p. 585-594.

Tsuji K. (2001). « Automatic extraction of translational japanese-katakana and english word pairs from bilingual corpora ». Proceedings of the 19th International Conference on Computer Processing of Oriental Langages (ICCPOL'01). p. 245-250.

Tsuji K. *et al*. (2002). « Extracting french-japanese word pairs from bilingual corpora based on translitération rules ». Proceedings of the third International Conference on Language Ressources and Evaluation (LREC 2002).

NAISSANCE ET CIRCULATION D'UN TERME : UNE HISTOIRE D'EXOPLANÈTES

Cristina Nicolae et Valérie Delavigne
Laboratoire LiDiFra - Université de Rouen

RÉSUMÉ

Dès lors que l'on s'intéresse à la circulation des termes dans différents types de discours, on s'aperçoit combien les formes terminologiques mobilisées s'y actualisent diversement. Elles varient selon les échanges, les acteurs, les supports, les lieux où elles circulent, les pratiques qui les convoquent et les moments où elles apparaissent. Nous engagerons ici notre parcours sur un sentier qui reste encore peu fréquenté : celui de la diachronie. A partir d'un ensemble de discours portant sur l'astrophysique, nous nous pencherons plus particulièrement sur le berceau d'un terme, « exoplanète », depuis sa naissance jusqu'à son essor dans les médias.

1 INTRODUCTION

Dès lors que l'on s'intéresse à la circulation des termes[33] dans différents types de discours, on s'aperçoit combien les formes terminologiques mobilisées s'y actualisent diversement. Elles varient selon les échanges, les acteurs, les supports, les lieux où elles circulent, les pratiques qui les convoquent et les moments où elles apparaissent. C'est sous ce dernier angle que nous engagerons ici notre parcours : celui de la diachronie, sentier qui reste encore peu fréquenté. Nous nous proposons d'illustrer cette circulation dans le flux de productions discursives particulières : à partir d'un ensemble de discours relatifs à l'astrophysique, nous nous pencherons plus particulièrement sur le berceau d'un terme, « exoplanète », depuis sa naissance jusqu'à son essor dans les médias.

En terminologie, les études diachroniques restent ponctuelles et marginales et ce, malgré l'intérêt de ces approches tout à la fois pour la description du fonctionnement sémantique des termes et, comme le souligne Pascaline Dury (2006), de façon générale pour l'épistémologie de la terminologie. Or cette dimension diachronique est centrale pour comprendre la façon dont un vocabulaire s'est constitué et évolue.

> L'étroite interdépendance de la langue et du discours fait qu'on ne saurait se passer d'un examen historique attentif des discours pour une juste description des terminologies. (Delavigne, 2006)

L'astrophysique est un domaine en plein bouleversement. Depuis une vingtaine d'années, le rythme dense des évolutions scientifiques ébranle l'usage et le sens des termes. Ce domaine constitue dès lors un champ d'observation remarquable et offre une belle opportunité pour

[33] Nous définissons le terme comme une unité lexicale dont la spécificité est à relier à son statut dans une communauté discursive donnée. Ce statut se manifeste dans le discours par des marques repérables (énoncés définitionnels, reformulations, connotations autonymiques, thématisations, etc.). Le terme ne devient tel que par décision du locuteur ou de l'analyste, qui le juge pertinent pour un savoir, un système de connaissances ou une pratique.

construire un corpus approprié à une analyse diachronique, procurant l'occasion d'examiner de près l'émergence, la circulation et l'évolution sémantique de termes sur une courte période.

L'astronomie a fait l'objet d'études linguistiques dans une perspective d'analyse de discours : Beacco (2000) visait ainsi à décrire les pratiques des médias dans la diffusion de savoirs complexes. Notre approche diffère ici dans la mesure où nous visons une description diachronique des terminologies en émergence dans ce domaine en évolution.

Notre article se propose de repérer l'actualisation des premières occurrences[34] du terme « exoplanète » dans des discours dits « de vulgarisation ». Nous comparerons les premiers contextes d'émergence de ce nouveau terme, puis en repèrerons le paradigme désignatif. Par la suite, nous confronterons quantitativement les différentes désignations dans les discours scientifiques et dans les discours de vulgarisation. Enfin, nous discuterons des différents usages en fonction des publications où elles apparaissent. En entrant dans le corpus par le bais du terme pivot « exoplanète », ce parcours constitue une première approche diachronique de ce domaine dynamique.

2 CONSTITUTION D'UN CORPUS DIACHRONIQUE

La première exoplanète est découverte en 1995 par Michel Mayor et Didier Queloz, astrophysiciens à l'Observatoire de Genève. Pour l'astrophysique, cette date est centrale car à partir de ce moment-là, le sujet se médiatise fortement aussi bien à l'intérieur qu'à l'extérieur de la communauté scientifique et les financements se font plus conséquents.

Le terme est enregistré dans le *Petit Robert* dès 1998. Il en propose la définition suivante :

> Exoplanète, n.f. – 1998; calque de l'anglais *exoplanete* (1996), de *extrasolar planet* « planète extrasolaire -> *exo-* et *planète*. (ASTRON.) planète orbitant autour d'une étoile qui n'appartient pas au système solaire (*Petit Robert*, 2002)

Le *Petit Robert* précise que la première attestation du terme anglais date de 1996. La forme « exoplanète » est présentée dans cette définition comme quasi-synonyme de « planète extrasolaire », dénomination qui était en usage avant l'émergence du signifiant *exoplanète*.

Une exoplanète est donc une planète qui se trouve en dehors de notre système solaire et qui est en orbite autour d'une autre étoile que le soleil. En partant à la recherche de ces astres, les spécialistes avaient comme unique modèle notre propre système solaire ; c'est pour cette raison que les découvertes n'ont cessé de surprendre car il s'avère que les systèmes planétaires découverts ont très peu de points en commun avec notre système solaire.

Si l'on considère la date de la découverte de la première exoplanète, on ne peut que souligner l'intégration très rapide du terme dans le dictionnaire, sans doute en lien avec la forte médiatisation du sujet. Notons de fait que le terme n'est pas sans rencontrer un certain succès. Deux exemples peuvent paraitre significatifs à cet égard : en 2008, un groupe de musique, Vinc2, sort un album intitulé *Exoplanète*[35] et en 2009, année mondiale de l'astrophysique, une série de timbres ayant pour titre *L'Astronomie* est éditée ; l'un d'entre eux s'intitule *exoplanète.*

[34] Nous employons le terme « occurrence » dans le sens de présence effective dans le discours d'une forme lexicale

[35] http://vinc2creations.free.fr/vinc2creations/index.php?page=articles&xa=1&xb=2

Tous les termes scientifiques ne connaissent pas cet accueil cordial dans la langue quotidienne.

2.1 Un corpus contrasté

Nous avons construit notre corpus autour d'un sous-domaine de l'astrophysique. Ce sous-domaine spécifique étudie la formation des planètes et recherche des planètes extrasolaires. Afin de contraster les usages, ce corpus prend en compte différents discours. Un ensemble d'articles réunis sous le nom de « corpus de vulgarisation » s'oppose aux articles dits « primaires »[36], écrits par des scientifiques à destination de leurs pairs et qui forment ce que nous appelons ici le « corpus scientifique ». Ce corpus scientifique se distingue du « corpus des communiqués de presse » du CNRS et du « corpus de vulgarisation » des revues *La Recherche* et *Science &Vie*, qu'on peut placer sous le syntagme de « discours de transmission des connaissances », pour reprendre le terme de Beacco et Moirand (1995).

Nous avons choisi d'examiner *La Recherche* et *Science &Vie* dans la mesure où ces deux revues sont considérées comme des productions « prototypiques » de vulgarisation, même si les deux publications ne visent pas le même public : *La Recherche* s'adresse plutôt à un public de chercheurs, d'enseignants et d'étudiants, tandis que *Science & Vie* vise un public plus jeune, sans formation spécialisée. On peut discuter du statut de la revue *La Recherche* qui se rapproche de ce que Louis Guespin désignait par « discours d'interface » (1991 : 74). Comme le signale Daniel Jacobi (1986), *La Recherche* pose un certain nombre de problèmes aux descripteurs de la vulgarisation, les chercheurs qui y proposent un article ayant le sentiment d'écrire « d'abord pour leurs pairs de la communauté scientifique » (Jacobi, 1986 : 41). Néanmoins, même si la revue entre plutôt dans la catégorie « semi-vulgarisation » pour utiliser la classification d'Anne-Marie Loffler-Laurian (1986), nous ne pouvons lui ôter son statut d'organe de vulgarisation dans la mesure où, comme le remarque Daniel Jacobi, les rédacteurs élaborent leurs articles également « pour un cercle beaucoup plus élargi que leurs lecteurs habituels » (Jacobi, 1986 : 42).

Les articles issus de la revue *Science & Vie* comptent 46 369 occurrences ; pour la revue *La Recherche*, nous en avons dénombré 46 146. Ces sous-corpus sont donc très homogènes d'un point de vue numérique et, dès lors, se prêtent à des observations quantitatives. Le sous-corpus formé des communiqués de presse du CNRS recueillis sur le site de cette institution couvre la période de 2002 à 2009 et compte pour sa part 13 368 occurrences. Nous avons ici des productions discursives émises sur un site labellisé scientifique, destinés par définition aux professionnels de la presse.

Le sous-corpus scientifique (les articles dits « primaires ») comprend pour sa part des articles portant sur les exoplanètes, produits lors d'une journée d'études française. Soulignons ici la rareté de ces documents ; en effet, depuis les années 1970, les publications dans ce

[36] Pour une discussion sur la notion de « discours de vulgarisation » et « discours primaire », voir Delavigne, 2001.

domaine se font en anglais, à quelques exceptions près. Les textes de notre corpus sont issus de l'une de ces rares situations où des chercheurs réunis en France étaient majoritairement francophones. Ces communications sont réunies sous le titre *Formation planétaire et exoplanètes*, document publié à l'occasion de l'École de Goutelas Astronomie en 2005[37]. Ce sous-corpus compte 38 612 mots, en provenance de cinq locuteurs différents. Afin d'accéder aux discours scientifiques et pallier ce qui forme un réel écueil méthodologique, un corpus d'entretiens doit être constitué.

2.2 Un corpus borné sur une diachronie courte

Nous prenons le terme de « diachronie courte » dans l'acception que lui donne Aurélie Picton. Elle souligne le fait qu'en terminologie, nous sommes souvent amenés à nous intéresser à des changements en cours :

> Ceci implique de fait d'adopter un point de vue dit « en diachronie courte » (Kyto*, et al.*, 2000). Cette nouvelle conception diachronique « courte » est nommée par Mair *brachychrony* (1997, repris dans Renouf, 2002) qui désigne ainsi l'étude du changement sur des intervalles courts de 10 à 30 ans.
>
> Bien que développée en langue générale, cette notion de diachronie courte est particulièrement pertinente en langue de spécialité où les changements observés sont en lien avec l'évolution d'un domaine scientifique, évolution souvent très rapide. (Picton, 2009 : 17)

L'année de découverte de la première exoplanète, 1995, a été constituée en date pivot de cette diachronie courte. Parcourant les deux revues *Science & Vie* et *La Recherche* sur une période de dix ans précédant l'année de la découverte de la première exoplanète, nous avons cherché à savoir si le sujet des planètes extrasolaires y était abordé. En effet, la recherche d'une vie extraterrestre et de planètes en dehors de notre système solaire fait partie des sujets repris régulièrement par les discours de vulgarisation. Avant même la découverte de la première exoplanète, dans le domaine de l'astrophysique, la recherche de planètes extrasolaires est un sujet régulièrement abordé : plusieurs découvertes ont d'ores et déjà été annoncées, bien que désavouées par la suite.

Il est intéressant de signaler que, dans ces deux revues, les articles portant sur ce sujet sont ponctuels avant 1995, alors que le sujet est *a priori* un des plus porteurs pour la vulgarisation. Des recherches dans d'autres publications ultérieures de vulgarisation et dans des journaux non spécialisés dans la médiatisation de l'information scientifique (comme *Le Monde,* par exemple) devraient nous permettre de comprendre mieux pourquoi un sujet aussi engageant que celui de la recherche de la vie ailleurs, thème indissociable de la recherche des exoplanètes, a mis si longtemps à rejoindre la grande famille des thématiques favorites de la vulgarisation.

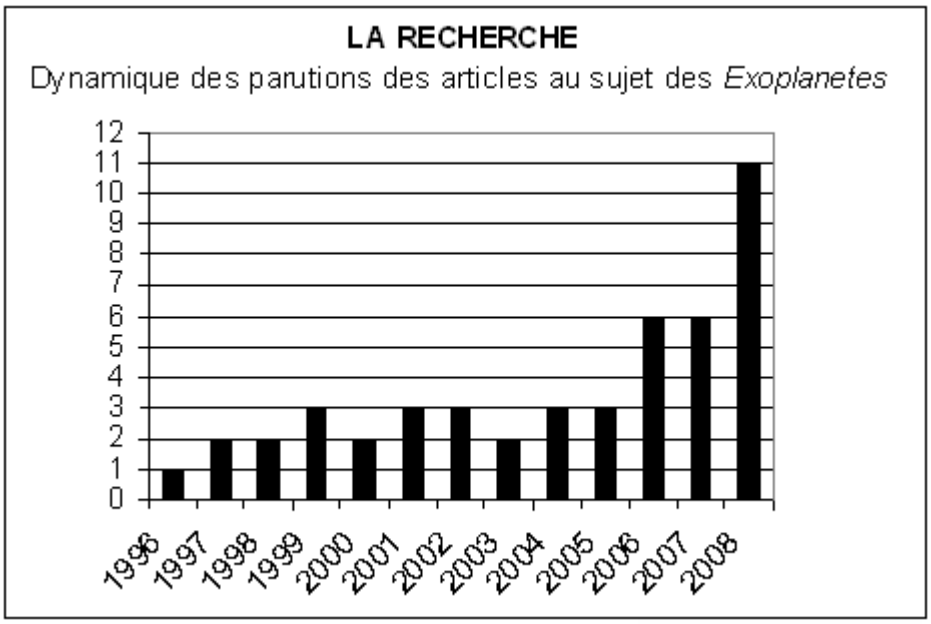
L'ensemble des articles portant sur les exoplanètes a été recensé de 1995 à 2009. Sont donc disponibles à l'observation des articles sur une période de vingt-quatre ans.

Pour une étude diachronique, cette période peut sembler très courte au premier abord. Néanmoins, comme il s'agit, rappelons-le, d'un domaine en plein bouleversement scientifique dont les connaissances sont en remaniement constant en fonction des nouvelles découvertes qui se succèdent à un rythme accéléré, cette diachronie courte est riche et pertinente pour l'analyse sémantique de certains des termes qui y sont actualisés.

Le tableau ci-dessus rend compte de la dynamique de parution des articles portant sur les exoplanètes et concernant la revue *La Recherche* (ce mouvement est moins marqué pour la

[37] Nous en profitons pour remercier M. Jean Schneider, astronome à l'Observatoire Paris-Meudon, qui a fourni ce matériel

revue *Science & Vie*). L'axe des abscisses, sur une échelle de 0 à 12, représente le nombre de fois où le sujet a été traité dans l'année, notée sur l'axe des ordonnées ; par exemple, en 2006, six articles portant sur les exoplanètes paraissent dans la revue.



Avant 1996, seules des parutions ponctuelles apparaissent sur le thème. A partir de la découverte de la première planète extrasolaire, nous observons un intérêt constant pour ce sujet ; on remarque la présence d'un article par an en 1996 pour passer à onze articles annuels en 2008, soit en moyenne un article chaque mois. On est donc passé d'un sujet marginal jusqu'en 1996 à une thématique constante de la revue.

3 PREMIÈRES OCCURRENCES DU TERME « EXOPLANÈTE » DANS LES DISCOURS DE VULGARISATION

3.1 Les discours de vulgarisation

Dans la revue *Science & Vie*, le terme « exoplanète » apparaît la première fois en janvier 2000, dans un article intitulé « On a enfin vu une planète extrasolaire » :

> En 1997, l'équipe de l'observatoire de Genève dirigée par Michel Mayor - qui a détecté, il y a quatre ans, la première **exoplanète**[38] autour d'une étoile -, se met sur la piste de l'étoile et tente d'y détecter la présence d'une exoplanète en son orbite. (*Science & Vie,* janvier 2002, « On a enfin vu une planète extrasolaire », Hervé Poirier)

Nous observons que cette première occurrence du nouveau terme « exoplanète » n'est nullement marquée et n'est accompagnée d'aucune définition ni d'aucune paraphrase de reformulation. Si nous examinons le paragraphe dans son entier, nous sommes devant un récit structuré, avec un ancrage temporel et une localisation précise. Le cadre pose sans ambiguïté l'observation des étoiles « en dehors de notre système solaire ».

[38] C'est nous qui marquons le terme de gras afin de faciliter au lecteur le repérage du terme et ce, dans l'ensemble des cotextes qui suivent.

Énumérons rapidement les informations issues de ce contexte et qui vont permettre le décodage de cette première occurrence d'« exoplanète ».

9. Un temps précis : « Mercredi 8 septembre 1999, 23h13, heure locale » ;
10. Un endroit précis : « Le ciel est bien dégagé sur les montagnes du Colorado (États-Unis) » ;
11. Un objet d'observation : l'étoile HD 209458, astre dans la constellation de Pégase, donc un astre en dehors de notre système solaire ;
12. Un objet recherché : « la silhouette d'une planète » supposée tourner autour de l'étoile.

Les éléments que l'on peut retenir tiennent dans le fait que l'étoile observée est située dans la constellation de Pégase et que les objets recherchées sont dénommés « planète » (trois fois) et « planète en dehors de notre système solaire » (une fois). Le terme « exoplanète » est « glissé » dans ce paradigme sans autre explication ; le lecteur doit déduire que le terme « exoplanète » entre dans la série :

40. Planète ;
41. Planète en dehors du système solaire.

On peut imaginer que pour l'auteur, le terme « exoplanète » est suffisamment transparent pour qu'il ne nécessite pas d'être défini : « la première exoplanète autour d'une étoile » est posé en équivalent de « la première *planète* autour d'une étoile » et, par inférence, le lecteur doit pouvoir conclure qu'« une exoplanète est une planète ».

Dans la revue *La Recherche,* la première occurrence du terme « exoplanète » apparaît dans un article de décembre 2002 : « Objectif Terres : mille et une méthodes de détection ».

> À seulement 6 années-lumière, mais trop peu brillante pour être vue à l'œil nu, cette étoile (l'étoile de Barnard) est la plus proche de notre Soleil, hormis le système triple Alpha du Centaure. Mais les observations des années soixante-dix démentirent cette première détection. D'autres annonces d'une planète extrasolaire, dite **exoplanète**, firent la « une » des médias, mais celles-ci non plus ne résistèrent pas à un examen minutieux. Ce n'est qu'après des dizaines de fausses découvertes que, en 1991, deux planètes extrasolaires furent réellement détectées. (*La Recherche,* décembre 2002, Jack J. Lissauer, « Objectif Terres : mille et une méthodes de détection »
>
> http://www.larecherche.fr/content/recherche/article?id=5784)

Le marqueur de reformulation « dite » indique un rapport de synonymie entre la désignation « planète extrasolaire » et le terme « planète ». Nous remarquons un certain degré de méfiance dans ce premier emploi par le choix d'un reformulant qui marque la distance entre le discours de l'auteur et cette nouvelle dénomination qu'il semble ne pas prendre en charge.

3.2 Les communiqués de presse du CNRS

Dans les communiqués de presse du CNRS, c'est en avril 2004 que le terme « exoplanète » apparaît pour la première fois :

> Ces processus d'évaporation complexes conduisent à des pertes de masse substantielles pour la planète qui affectent directement sa structure - donc son rayon - et son évolution, voire sa durée de vie. Les calculs de l'équipe du CRAL mettent en évidence le concept de masse critique pour ces **exoplanètes** fortement irradiées, variant avec la distance orbitale. (Paris, 21 avril 2004, « Les planètes extrasolaires : des modèles pour comprendre leurs évolutions », http://www2.cnrs.fr/presse/communique/461.htm)

Signalons ici que les communiqués de presse du CNRS portant sur ce thème ne font leur apparition qu'en 2002. C'est pourtant pour la communauté scientifique un sujet de réflexion

depuis les années soixante-dix. Cette donnée contraste avec le fait que, dans les revues de vulgarisation, le thème est constamment abordé depuis 1996.

Le terme « exoplanète » apparaît dans ce sous-corpus pour la première fois en 2004. Comme dans la revue *Science & Vie*, il n'est accompagné dans le corps de l'article d'aucune définition, ni même reformulé. En revanche, à la fin de l'article, un encadré propose un point sur les exoplanètes. Nous y observons une reformulation d'« exoplanète » avec le marqueur « dite ». Le qualificatif « extrasolaire » qui accompagne « exoplanète » est également reformulé.

> Plus de cent planètes gazeuses ont été découvertes à ce jour en orbite autour d'étoiles autres que notre Soleil. Le domaine de recherche des **planètes dites extrasolaires ou exoplanètes** représente de fait un pan nouveau de l'astronomie du 21e siècle, au confluent de domaines tels que la planétologie, l'astrophysique stellaire et la physique de la matière dense. (Paris, 21 avril 2004, « Les planètes extrasolaires : des modèles pour comprendre leurs évolutions », http://www2.cnrs.fr/presse/communique/461.htm)

Le rapport de synonymie s'établit facilement, mais la mise à distance du qualificatif « extrasolaire » peut étonner dans la mesure où cette dénomination, « planète extrasolaire », était la plus utilisée dans les discours scientifiques et de vulgarisation avant l'apparition du terme « exoplanète ».

Dans les trois cas observés ci-dessus, il nous faut donc souligner l'absence de définition, que ce soit à l'intérieur des articles proprement dits ou dans la supra-structure des articles. Cette situation n'est pas rare en vulgarisation scientifique, mais pour un terme dont le formant *exo-*, d'origine grecque, qui correspond à la locution prépositive « hors de », « à l'extérieur de » (cf. Guilbert, 1971 : XLVIII), est spécifique des vocabulaires scientifiques, il semble légitime de s'attendre à plus de précision, particulièrement pour les premiers emplois. Les contextes proches et les reformulations avec le marqueur « dite » créent la possibilité d'établir une relation d'hyperonymie entre les termes « planète » et « exoplanète », tout comme la relation synonymique entre « planète extrasolaire » et « exoplanète » facilite la relation d'inclusion entre « planète » et « planète extrasolaire ».

4 PARADIGME DÉSIGNATIF DU TERME « EXOPLANÈTE » : COMPARAISON QUANTITATIVE ENTRE LES DIFFÉRENTS SOUS-CORPUS

Pour étudier les différentes dénominations et désignations[39] autour du terme « exoplanète », nous empruntons le concept de « paradigme désignatif » à Marie-Françoise Mortureux qui le définit comme une « liste de syntagmes (en général nominaux parfois verbaux) fonctionnant en coréférence avec un vocable initial dans un discours donné » (1993 :124). Celui-ci nous permet d'observer certains des moyens discursifs de dénomination des exoplanètes et le rapport entre « exoplanète » et ses autres désignations.

A partir de l'ensemble du corpus, nous avons pu construire le paradigme désignatif suivant :

11. exoplanète
12. planète extrasolaire
13. planète extrasolaire géante
14. planète extérieure (à notre système solaire)
15. planète géante
16. géante gazeuse

[39] Pour la différence « dénomination » vs « désignation », voir Kleiber, 1984.

17. Jupiter chaud
18. Super-Terre
19. planète

Le tableau présente la distribution de ces désignations par sous-corpus.

Désignation	CNRS (13 368 mots)	Science & Vie (46 369 mots)	La Recherche (46 146 mots)	Corpus scientifique (38 612 mots)
exoplanète	41	101	80	27
planète extrasolaire	37	27	78	32
planète extrasolaire géante	1	1	0	0
planète extérieure (à notre système solaire)	1	5	0	0
planète géante	17	18	104	15
géante gazeuse	3, dont 2 « planètes géantes gazeuses »	17, dont 1 « planète géante gazeuse »	7, dont 5 « planète géante gazeuse »	1
Jupiter chaud	1, avec guillemets	1, avec guillemets	5, dont 4 avec guillemets « Jupiter chaud »	16, dont 3 avec guillemets
super-Terre	1 sans guillemets	14, dont 6 avec guillemets	3 sans guillemets	1, sans guillemets
planète (avec différents qualificatifs, y compris ceux du tableau)	184	749	808	305

Arrêterons-nous sur les présences significatives, ainsi que les différences et les ressemblances entre les différentes désignations, qui font apparaître des caractéristiques propres à chaque publication.

4.1 Exoplanète/planète extrasolaire

Un des plus importants écarts à remarquer est au niveau numérique de l'usage du terme « exoplanète » : là où la vulgarisation convoque 101 occurrences pour *Science & Vie* et 80 occurrences pour *La Recherche* (rappelons que ces deux sous-corpus sont de taille homogène), nous repérons seulement 27 occurrences dans le corpus scientifique. En revanche,

les 41 occurrences du terme « exoplanète » dans le sous-corpus des communiqués de presse CNRS restent un chiffre important si on le ramène aux dimensions de ce sous-corpus.

Nous constatons une préférence pour ce nouveau terme dans le corpus de vulgarisation, par opposition aux discours scientifiques ; malgré la faible circulation des termes formés avec le préfixe *exo-*, la connotation scientifique du terme et sa transparence semblent plaider pour son utilisation préférée dans les discours de vulgarisation.

Si nous comparons l'emploi du terme « exoplanète » avec la désignation la plus utilisée avant sa mise en circulation, « planète extrasolaire », des différences importantes apparaissent entre les deux publications de vulgarisation. Nous pouvons en comptabiliser 27 occurrences seulement dans la revue *Science & Vie* (pour 101 occurrences d' « exoplanète ») et 78 dans la revue *La Recherche*, qui maintient ainsi un parfait équilibre par rapport aux 80 occurrences du terme « exoplanète ». Les mêmes proportions équilibrées sont à remarquer dans le corpus scientifique, avec 32 occurrences du terme « planète extrasolaire » pour 28 occurrences du terme « exoplanète ».

Cet usage des termes « planète extrasolaire » et « exoplanète » confirme le rapprochement entre les discours scientifiques et les discours de la revue *La Recherche*, généralement considérée comme publication de semi-vulgarisation, à mi-chemin entre la vulgarisation et les discours scientifiques (Delavigne, 2001). En même temps, pour une publication de vulgarisation dont le public est constitué essentiellement par des lycéens et des amateurs de science (conformément aux enquêtes de Daniel Jacobi (1999) sur les publics-cible de ces deux revues) comme la revue *Science & Vie*, la circulation plus large du nouveau terme ne surprend guère, vu sa transparence et sa connotation « scientifique ».

4.2 Planète géante

Les planètes géantes sont des planètes gazeuses, d'où leur désignation concurrente « géantes gazeuses ». Ce sont des planètes dites « de type Jupiter » qui représentent une sous-catégorie de planètes. La majorité des exoplanètes découvertes à ce jour fait partie de cette classe de « planètes géantes » ou de « Jupiters chauds ». Les moyens d'observation dont disposent aujourd'hui les astrophysiciens n'ont permis, jusqu'à présent (fin 2009, date de la borne supérieure de notre corpus), de découvrir que quatre planètes rocheuses du même type que la Terre.

« Planète géante » est une désignation employée massivement dans la publication *La Recherche* : 104 occurrences pour 15 seulement en discours scientifique et 18 dans la revue *Science & Vie*. Cette dénomination est bien représentée, sans doute parce qu'elle est un descripteur transparent par rapport à son hyperonyme « planète ».

Notre recherche étant en cours, nous ne disposons pas assez d'éléments actuellement pour interpréter cette forte différence quantitative entre les deux revues *Science & Vie* et *La Recherche*. Nous pourrions l'expliquer de la même façon que précédemment : nous serions en présence d'un discours plus proche du discours « primaire » pour la revue *La Recherche*, et d'un discours plus vulgarisé en ce qui concerne *Science & Vie*.

4.3 Planète et exoplanète

Précisons ici que nous n'avons pas dégroupé les termes composés, du type « planète extrasolaire » des termes simples, du type « planète ». Le nombre d'occurrences du terme « planète » est donc inclus dans celui des termes composés comme « planète géante » ou « planète extrasolaire ». Avant d'affiner ultérieurement ces observations, constatons seulement l'usage plus fréquent du terme « planète » en tant que générique incluant la sous-

classe « exoplanète », avec la prise en charge du chevauchement terme simple-terme complexe.

Les « exoplanètes » étant tout d'abord des *planètes*[40], une fois posé le cadre d'observation des étoiles et des planètes *en dehors du système solaire*, il n'est plus nécessaire d'utiliser systématiquement le terme « exoplanète » : « planète » suffit, en vertu de la relation d'hyperonymie. C'est ce qui explique l'importante disproportion entre le nombre d'occurrences de « planète » et d'« exoplanète ».

On peut même aller jusqu'à se demander quelle sera la longévité du terme « exoplanète » si fortement concurrencé par des signes du même champ lexical.

4.4 Jupiter chaud, super-Terre

Deux termes ont pour caractéristique commune leur dimension métaphorique : « Jupiter chaud » et « super-Terre ». Cette situation est fréquente dans les dénominations de notions scientifiques. De cette dimension métaphorique découle une instabilité qui se manifeste aussi bien au niveau graphique, par la présence/absence de tiret (« Jupiter-chaud »/« Jupiter chaud », « super Terre »/« super-Terre »), que dans la prise en charge de ces dénominations, traduite par la présence/absence de guillemets (Jupiter chaud/« Jupiter chaud », super Terre/« super Terre ») ; ces signes typographiques jouent sur l'interprétation que peut en faire le lecteur. Cette instabilité est à signaler pour l'ensemble du corpus. La question qui se pose est de savoir si nous sommes devant des termes désignant des sous-catégories d'exoplanètes ou si nous avons affaire à de simples analogies auxquelles les auteurs font appel pour faciliter la compréhension des notions.

4.4.1 JUPITER CHAUD

Dans les contextes qui suivent, nous observons que dans les discours de vulgarisation, cette désignation n'est pas toujours prise en charge, les auteurs renvoyant aux « experts » - les astronomes - conformément aux usages de la vulgarisation aujourd'hui (cf. Reboul-Touré, 2004, par exemple), soit par l'utilisation de guillemets, soit par des formules définitionnelles du type « être appelé », voire les deux.

> Mais depuis octobre 1995, quelque 200 nouvelles planètes ont été détectées autour d'autres étoiles que le Soleil, la plupart avec des caractéristiques orbitales si curieuses qu'elles prennent en défaut ce scénario. Les astronomes ont en effet découvert ce qu'ils appellent des « **Jupiters chauds** » : d'énormes boules de gaz tournant autour de leur étoile... à quelques centièmes d'UA en quelques jours à peine ! (*Science & Vie*, oct. 2006, pp. 130-131, « Questions-réponses »)

> D'autres planètes similaires ont été détectées ultérieurement. Cela veut dire qu'environ 1 % des étoiles de type solaire possèdent des planètes de la taille de Jupiter dont la période orbitale ne dépasse pas une semaine. Grande masse et orbite proche de l'étoile, on les appelle des « **Jupiters chauds** ». (*La Recherche,* décembre 2002, « Objectif Terres : mille et une méthodes de détection », Jack J. Lissauer)

> Ce « **Jupiter chaud** » tourne en 3,35 jours autour de HD 188753 A, une étoile de type solaire, qui tourne elle-même en plus de vingt-cinq ans autour d'un couple d'étoiles. (*La recherche,* Septembre 2005, « 162 exoplanètes »).

Il faut noter la présence de guillemets autour de la dénomination « Jupiter chaud » dans les discours de vulgarisation comme dans les discours scientifique. Marques de citation ou signe

[40] Un travail est en cours sur la circularité des définitions des termes fournis par les experts en discours de vulgarisation et sur l'insuffisance du modèle des conditions nécessaires et suffisantes (CNS) à rendre compte du fonctionnement sémantique des termes appartenant à des domaines en plein bouleversement comme celui qui nous occupe ici.

de distance ? Les deux hypothèses sont possibles. C'est un des faits qui conduit à poursuivre nos recherches sur le discours définitoire dans le discours des experts astrophysiciens vers le grand public.

Qui plus est, dans le discours scientifique, nous retrouvons en guise de formule définitoire le verbe « surnommer », ce qui peut laisser entendre que la dénomination *Jupiter chaud* n'est pas un terme :

> La découverte de nombreuses planètes extra-solaires par la méthode des vitesses radiales (Mayor & Queloz, 1995), en particulier celles qui orbitent très près de leur étoile et sont surnommées les « **Jupiters chauds** », a fait espérer à de nombreux observateurs que l'une d'entre elles serait bientôt découverte en train de passer devant le disque de son étoile. (Alfred-Vidal Madjar, *L'évaporation de la planète Osiris*, Journées d'étude, Goutelas, 2005)

Surnom, métaphore ou terme scientifique ? Il est difficile de décider du statut de cette désignation à partir seulement des contextes du corpus. Pour ce faire, il faut consulter d'autres sources. Si l'on consulte le site le l'Observatoire de Paris-Meudon[41] sur lequel des scientifiques diffusent des connaissances pour le grand public, il apparaît que les Jupiters chauds sont bel et bien une sous-catégorie de planètes extrasolaires.

4.4.2 SUPER-TERRE

Nous sommes avec « super-Terre » dans le même cas de figure. La seule occurrence dans les discours scientifiques, sans guillemets, oriente l'interprétation vers une dénomination rigoureuse d'une sous-classe d'exoplanètes ; cependant, l'absence de définition de cette sous-catégorie d'astres, à laquelle s'ajoute la difficulté d'une généralisation à partir d'un seul exemple, nécessitent de prendre certaines précautions avant de se prononcer sur le statut de cette désignation : terme ou figure de style ?

Comme nous pouvons le constater dans l'exemple ci-dessous, il n'est guère facile de tirer de conclusion :

> On peut raisonnablement penser que, ces prochaines années, les mesures Doppler franchiront la barre du km h^{-1} (28 cm s^{-1}) permettant alors de détecter des **Terres chaudes**, et même des **super-Terres** dans la zone habitable. (F. Bouchy, « Détection des planètes par méthodes radiales », École de Goutelas, 2005)

L'interprétation la plus plausible, liée au statut du discours, de type scientifique, est que « super-Terre » est un terme qui désigne une sous-catégorie d'exoplanètes. Cette interprétation de « super-Terre » comme unité terminologique est renforcée par l'absence de guillemets dans les corpus de la revue *La Recherche.* Certaines explications et reformulations permettent de dégager les caractéristiques de cette sous-classe. Les super-Terres sont des planètes extrasolaires de type rocheux, avoisinant la masse de la Terre et plus susceptibles d'abriter la vie que les géantes gazeuses.

Convoquons ci-dessous deux exemples de la revue *La Recherche* :

> Une quatrième **super-Terre**, exoplanète rocheuse avoisinant la masse de la Terre (entre une et dix masses terriennes) aurait été repérée. (*La Recherche*, « Exoplanète et méthane », mai 2008, http://www.larecherche.fr/content/recherche/article?id=23215)
>
> L'un des graal de la recherche d'exoplanètes est de découvrir des petites planètes rocheuses, du type de la Terre, ailleurs dans l'Univers. Parmi les 330 connues à ce jour, on compte une vingtaine de **super-Terres**, c'est-à-dire des planètes dont la masse ne dépasse pas 15 masses terrestres. […] C'est donc la première fois que l'on mesure la taille d'une **super-Terre** et son rayon vaut 1,75 fois celui de la Terre ! (*La Recherche,*

[41] http://media4.obspm.fr/exoplanetes/pages_definition/questce.html

février 2009, *La plus petite exoplanète jamais mesurée,* Hélène le Meur, *http://www.larecherche.fr/content/recherche/article?id=24873*)

Cependant, la confrontation des sous-corpus nous laisse hésiter sur le statut de cette désignation. Dans *Science & Vie,* trois sur les seize occurrences sont guillemetées. Signes de citation ou de mise à distance ? Comme dans le cas de « Jupiter chaud », l'interprétation de cette désignation en tant que terme nécessite l'appel à d'autres éléments.

Citons deux exemples issus de *Science & Vie*, dont le premier est un sous-titre :

> Voici peut-être la première « **super Terre** ». (*Science & Vie*, Octobre 2004, « On a découvert une autre terre », Laurent Orluc, Benjamin Noël, Valérie Greffoz)
>
> Et rien que dans les mois qui viennent, confie Jean Schneider, je ne serais pas étonné que l'on trouve des « super-Terre », c'est-à-dire des planètes environ dix fois plus massives que la nôtre, par la méthode dite des vitesses radiales. (*Science & Vie*, Février 2003, « Exoplanètes : On est enfin tout près de les voir ! », Valérie Greffoz)

Le syntagme « super-Terre » apparaît ici guillemeté. L'expert convoqué en propose une reformulation en forme de définition, qui nous pousse vers l'hypothèse que « super-Terre » possède un statut terminologique. Cependant, l'usage qu'en fait la revue *Science & Vie*, qui joue sur les valeurs métaphoriques du signe, instaure un doute : métaphore ? Terme ? Comme dans le cas de « Jupiter chaud », l'absence de définition claire contraint dès lors à faire appel à d'autres sources pour décider du statut que peut prendre cette désignation.

5 CONCLUSION

Découvertes récentes, les planètes situées hors de notre système solaire sont désignées par le néologisme « exoplanète », apparu en français en 1998. Pistant le terme, nous avons pu constater son entrée rapide dans le dictionnaire, très proche du moment de la découverte de la première planète extrasolaire, ainsi que son accueil favorable hors du domaine de l'astrophysique. Notre étude a permis d'en repérer les premières occurrences dans les discours de vulgarisation (2000 pour la revue *Science & Vie*, 2002 pour la revue *La Recherche*).

Nous avons pu voir que l'émergence d'un nouveau terme n'est pas systématiquement accompagnée d'une définition claire dans les discours de vulgarisation et que les traces de « didacticité » (Beacco et Moirand, 1995) sont souvent absentes.

L'analyse du corpus a mis en évidence des différences quantitatives importantes dans l'usage des désignations en fonction des types de publication. Mais elle ne nous a pas permis de décider du statut, terminologique ou non, de certains désignatifs comme « Jupiter chaud » et « super-Terre » dans la mesure où les énonciateurs jouent sur la dimension métaphorique de ces unités. Pour lever l'ambigüité et décider de leur statut, il est alors nécessaire d'élargir le corpus et de consulter d'autres sources.

L'analyse permet de mettre au jour les caractéristiques des variantes désignatives des exoplanètes. L'étude quantitative entreprise révèle des variations graphiques de certaines désignations (comme « super Terre » ou « Jupiter chaud ») et les divers modes de prise en charge (« planète extrasolaire *dite* exoplanète », présence/absence de marqueurs de distance). Cette étude fournit un socle pour poursuivre l'analyse du champ lexical d'« exoplanète » : si elle montre l'intérêt de porter un regard sur la diachronie d'un domaine en mouvement, se tourner vers la sémantique des termes devrait ensuite permettre d'en saisir l'évolution et de mieux comprendre comment un champ sémantique se refaçonne dès lors qu'émerge la nouveauté.

6 RÉFÉRENCES

Beacco J.-C. (2000). *L'astronomie dans les médias*. Paris : Presses de la Sorbonne Nouvelle.

Beacco J.-C. et Moirand S. (1995). « Autour des discours de transmission de connaissances ». *Langages*, 117, p. 32-53.

Delavigne V. (2001). *Les mots du nucléaire. Contribution socioterminologique à une analyse des discours de vulgarisation.* Thèse de Doctorat en Sciences du Langage, Université de Rouen.

Delavigne V. (2006). « La formation du vocabulaire de la physique nucléaire : quelques jalons ». Dans F. Gaudin et D. Candel, *Aspects diachroniques du vocabulaire*. Rouen : Publications des Universités de Rouen et du Havre. p. 89-107.

Dury P. (2006). « La dimension diachronique en terminologie et en traduction spécialisée : le cas de l'écologie ». Dans F. Gaudin et D. Candel, *Aspects diachroniques du vocabulaire*, Rouen : Publications des Universités de Rouen et du Havre. p. 109-124.

Gaudin F. (1993). *Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles*. Rouen : Publications de l'Université de Rouen.

Guespin L. (1991). « La circulation terminologique et les rapports, science technique, production ». *Cahiers de linguistique sociale*, 18, p. 59-80.

Guilbert L. (1971). « De la formation des unités lexicales. Fondements lexicologiques du dictionnaire ». *Grand Larousse de la langue française* vol.1. Paris : Larousse. p. IX-LXXX.

Jacobi D. (1986). Diffusion et vulgarisation : itinéraires du texte scientifique. Paris : Les Belles lettres.

Jacobi D. (1999). *La communication scientifique : discours, figures, modèles.* Grenoble : Presses Universitaires de Grenoble.

Jeanneret Y. (1994). *Écrire la science. Formes et enjeux de la vulgarisation.* Paris : Presses Universitaires de France.

Kleiber G. (1984). « Dénomination et Relations Dénominatives ». *Langages*, 76, p. 77-9.

Loffler-Laurian A.-M. (1983). « Typologie des discours scientifiques : deux approches ». *Études de Linguistique appliquée*, 51, p. 8-20.

Moirand S. (1992). « Autour de la notion de didacticité ». Les Carnets du CEDISCOR, 1, p. 9-20.

Mortureux M.-F. (1993). « Paradigmes désignationnels ». Semen, 8, p. 123-141.

Picton A. (2009). *Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial*. Thèse de Doctorat en Sciences du Langage, Université de Toulouse.

Reboul-Touré S. (2004). « Écrire la vulgarisation scientifique aujourd'hui ». *Colloque Sciences, Médias et Société.* Lyon : ENS-LSH, http://sciences-medias.ens-lsh.fr/article.php3_?id_article=65

LE RÔLE DE LA PROSODIE DANS LE TRAITEMENT AUTOMATIQUE DU SENS : L'EXEMPLE DE *ENFIN* DANS UN CORPUS DE FRANÇAIS PARLÉ

Mélanie Petit
Laboratoire Ligérien de Linguistique

RÉSUMÉ

Nous proposons dans cet article de prendre en considération la dimension prosodique dans la caractérisation sémantico-pragmatique des connecteurs discursifs en français. Nous illustrerons nos propos à partir d'une étude de *enfin* en français parlé spontané. Nous définirons au cours de notre étude, et suite à la présentation de nos analyses et résultats, un format ternaire de représentation du sens qui apparaît comme une alternative à la distinction signification/sens et que nous pensons pouvoir servir d'appui au traitement automatique du sens.

1 INTRODUCTION

L'objectif de notre étude est d'intégrer la dimension prosodique dans la mise en place d'une discrimination de la polysémie des unités lexicales, autrement dit de définir un processus de discrimination prosodique du sens, afin notamment de montrer qu'une caractérisation sémantique des unités du lexique, ainsi affinée et reposant sur des critères objectifs que sont les paramètres prosodiques, pourrait participer à développer le traitement automatique du sens. Nous illustrerons ces hypothèses en proposant une analyse sémantico-pragmatique et prosodique du connecteur *enfin* en français parlé spontané.

2 OBJET D'ÉTUDE

Nous avons orienté nos recherches sur le connecteur *enfin* pour différentes raisons. Tout d'abord, d'un point de vue général, les connecteurs discursifs sont des unités fondamentales dans l'organisation et la structuration du discours. Pour cette raison, nous estimons que leur étude est susceptible de présenter un grand intérêt en termes de traitement automatique du sens. Par ailleurs, *enfin* s'avère être une unité très étudiée majoritairement dans une perspective de caractérisation sémantique (Cadiot & al., 1985 ; Franckel, 1987 ; Luscher & Moeschler, 1990 ; Hwang, 1993 ; Barnes, 1995 ; Rossari, 1997, 2000 et 2004 ; Nemo, 2000 ; Beeching, 2000, 2001, 2002 et 2007 ; Razgouliaeva, 2002 et 2004 ; Paillard, 2003; Chanet, 2003 ;; Hansen, 2005a et 2005b ; Waltereit, 2007 ou encore Buchi & Stadtler, 2008), mais également quelquefois sous l'angle de la prosodie (Fónagy, 1983 ; Bertrand & Chanet, 2005). De ce fait, cela nous a permis de disposer d'une quantité non négligeable d'études portant sur le même objet que le nôtre et qui nous ont servi d'appui dans la réalisation de nos travaux. Enfin, notre choix s'est précisément porté sur cet item car celui-ci a développé, comme nous allons le voir au cours de cette étude, une polysémie importante, nous laissant ainsi la possibilité de tester l'existence d'une discrimination prosodique du sens de manière approfondie.

3 ÉTAT DE LA QUESTION

Avant d'aborder plus à même la présentation de nos recherches, nous allons brièvement nous intéresser à l'état de la question de la prise en compte de la dimension prosodique dans les études relatives aux connecteurs et plus globalement aux mots de discours, en français. Il ressort de notre tour d'horizon de la littérature que la très grande majorité des descriptions sémantico-pragmatiques des mots de discours ne prennent pas en considération la prosodie. Il faut néanmoins signaler quelques travaux qui font exception à ce constat :

Bertrand & Chanet (2005) s'intéressent aux liens entre la prosodie des connecteurs *enfin* et *voilà* et leurs différents rôles dans le discours. Elles concluent à l'absence de lien entre un sens spécifique et une intonation.

Noda (2005) propose une analyse des emplois de « hein » dans l'organisation des rapports intersubjectifs en traitant notamment du positionnement de cette unité dans l'énoncé ainsi que de la prosodie avec laquelle elle est réalisée. Son travail l'amène à s'interroger sur la possibilité d'identifier la valeur sémantique de « hein » en se fondant sur l'analyse prosodique.

Léglise (1999) s'est également intéressée à la question de la mie au jour de liens stables entre différentes valeurs sémantiques de *hein* et des réalisations prosodiques particulières de cette unité.

Dostie (2004) propose pour sa part un traitement des marqueurs discursifs selon une double perspective, à savoir en termes de la relation entre la sémantique et la lexicologie, et du point de vue de leur degré de pragmaticalisation. Elle souligne l'intérêt de prendre en compte la prosodie pour repérer les divers sens d'un marqueur et opte pour une analyse perceptuelle de ce paramètre.

Vincent & Demers (1994) ont étudié la classe des ponctuants en cherchant plus précisément à en fournir une définition claire et à en préciser les caractéristiques prosodiques. Elles illustrent ce travail à propos de *là*.

G. Cloiseau (2007) a consacré son doctorat à l'étude du phénomène de la métaphoricité à l'oral à partir d'interviews en français et en anglais portant sur le thème de la musique. L'auteur propose des outils de détection de la métaphore à l'aide de balises morphosyntaxiques et sémantiques mais également par l'observation de patrons accentuels en soulignant l'intérêt d'un tel travail pour la traduction.

Ces travaux se situent généralement dans le domaine de la sémantique. Nous avons également pu observer que les chercheurs en prosodie considèrent pour leur part traditionnellement dans leurs études des unités relevant d'un niveau supra-lexical sans prise en compte *a fortiori* de la question de la polysémie lexicale. Un fort cloisonnement demeure en effet entre ces deux disciplines.

4 HYPOTHÈSE

A l'origine de notre étude, nous avons émis l'hypothèse, fondée sur l'intuition, selon laquelle il était manifeste que la prosodie jouait un rôle dans le fait que l'auditeur accédait

spontanément à la bonne interprétation d'une occurrence de *enfin* réalisée isolément, selon que celle-ci exprime l'irritation, le soulagement ou encore la résignation. C'est donc dans un premier temps à la vérification de cette hypothèse que nous allons consacrer notre étude.

5 PRÉSENTATION DU CORPUS

Notre corpus de travail est constitué de 199 occurrences orales de *enfin* comportant la plus large diversité d'emplois possible. Les locuteurs sont des hommes et des femmes âgés de plus de 18 ans et parlant un français standard. La base de données principale dans laquelle nous avons extrait nos données est l'*Enquête SocioLinguistique à Orléans* 1 (ESLO 1) qui présente l'avantage, d'une part de contenir plus de 300 heures d'enregistrement en français parlé spontané et d'autre part, de proposer une documentation précise des locuteurs et des situations d'enregistrement. Nos sources secondaires sont des émissions de radio, des magazines télévisuels, des films et des pièces de théâtre ainsi que des lectures de nouvelles. Figurent dans ce tableau le nombre d'occurrences extraites par base de données et le pourcentage que chaque base de données représente par rapport à la totalité de notre corpus :

BASE DE DONNÉES	**NOMBRE D'OCCURRENCES = POURCENTAGE**
***ENQUÊTE SOCIOLINGUISTIQUE À ORLÉANS* 1**	**88 = 44,5%**
DES SOUS ET DES HOMMES	**18 = 9%**
C'EST PAS SORCIER	**9 = 4,5%**
C'EST DANS L'AIR	**9 = 4,5%**
RIPOSTES	**2 = 1%**
FRANCE TÉLÉVISION	**24 = 12%**
FAITES ENTRER L'ACCUSÉ	**12 = 6%**
PIÈCES À CONVICTIONS	**1 = 0,5%**
DISCOURS DE NICOLAS SARKOZY	**4 = 2%**
SPECTACLE DE PIERRE DESPROGES	**3 = 1,5%**
QUELLE FAMILLE	**7 = 3,5**
DOM JUAN	**7 = 3,5**
TAIS-TOI	**1= 0,5%**
SE SOUVENIR DES BELLES CHOSES	**2 = 1%**
MISERY	**1 = 0,5%**
SHINING	**1= 0,5%**
LE DERNIER TRAPPEUR	**2 = 1%**
LECTURES DE PIERRE BELLEMARRE	**4 = 2%**
IN THE MOOD FOR LOVE	**2 = 1%**
ALIAS	**2 = 1%**
TOTAL	**199 = 100%**

La caractérisation prosodique des données a consisté, à l'aide du logiciel Praat (www.fon.hum.uva.nl/praat/), en l'étude des paramètres prosodiques suivants :

42. forme de Fo[42] sur [enfin]

[42] La Fo désigne la fréquence fondamentale qui permet aux auditeurs de percevoir les sons comme graves ou aigus. Nous parlerons également indifféremment de *mélodie*.

43. forme de Fo sur [en]
44. forme de Fo sur [fin]
45. localisation du maxima de Fo
46. observation d'une rupture ou d'une intégration prosodique de *enfin* par rapport à son contexte
47. longueur de [enfin]
48. longueur de [en]
49. longueur de [fin]
50. ratio de la longueur de [en] sur longueur de [fin]
51. observation de la courbe d'intensité

Par ailleurs la configuration prosodique des contextes droit et gauche a également été prise en considération, la réalisation du connecteur pouvant être influencée par la localisation syntaxique de celui-ci. Nous avons procédé de la sorte afin de définir le comportement prosodique de chaque occurrence de notre corpus de *enfin*.

Nous souhaitons apporter quelques précisions quant à la caractérisation prosodique que nous proposons. Ne figurent ici que les paramètres que nous avons considérés, après en avoir testés plusieurs, comme pertinents dans le cadre de notre étude. Une observation préalable de la prosodie de *enfin* relativement à la prosodie du segment textuel dans lequel il est inséré ne nous a pas permis de prédire systématiquement le comportement prosodique de *enfin*. Nous ajouterons que des paramètres habituellement considérés comme non pertinent tels que la configuration mélodique de la première syllabe se sont avérés pertinents dans le cadre de notre étude. Pour cette raison, nous les avons conservés. Ainsi, sans rejeter par ailleurs l'existence maintes fois évoquées par différents prosodistes d'un lien fort entretenu entre la prosodie et la syntaxe, nous allons tester l'existence simultanée d'une prosodie jouant un rôle dans la désambigüisation lexicale.

6 ANALYSE DE *ENFIN*

6.1 Détermination du point de départ

Afin de procéder concrètement à la mise au jour de paires « formes prosodiques/sens », nous nous sommes retrouvée dans un premier temps face à l'alternative consistant à prendre comme point de départ soit le sens, soit la dimension sonore. Nous avons adopté pour une méthodologie initiale consistant à partir d'un premier classement sémantique des emplois de *enfin*, notre recherche se situant avant tout dans le cadre d'une sémantique linguistique consistant à traiter de la diversité des emplois d'un signe.

6.2 Elaboration du classement sémantique

Une première approche consistant à faire correspondre les occurrences de notre corpus avec l'une ou l'autre des propositions de classement faites dans les travaux relatifs à *enfin* et cités précédemment ne nous a pas permis d'aboutir à un résultat concluant. Il s'avère en effet que les données recueillies dans des corpus authentiques sont plus riches et plus variées que la caractérisation sémantico-pragmatique qui en est faite dans la littérature. Nous inspirant néanmoins de ces recherches et considérant la signification morphémique (commune à tous les emplois) de *enfin* telle qu'elle a été définie par Nemo (2000) :

« Indication qu'un problème se pose à un moment t et est résolu en t+1 »,

nous avons élaboré notre propre classement de travail. Pour ce faire, il a s'agit de se poser la question de savoir quelle était la nature du problème (discursif ou situationnel) et à quel

moment était réalisé *enfin* par le locuteur relativement au moment où se posait le problème. Nous avons de cette manière défini ce que nous appellerons des *profils* (au sens de Cadiot & Visetti, 2001). Ces profils correspondent donc aux types d'emplois de notre classement. Voici quelques exemples de profilages de *enfin* :

- emploi de soulagement : le problème (situationnel) se posait en $\mathbf{t_{-2}}$ et a été résolu en $\mathbf{t_{-1}}$
- emploi d'irritation : le problème (situationnel) se pose en $\mathbf{t_0}$ et doit être résolu en $\mathbf{t_{+1}}$
- emploi de résignation : le problème (situationnel) se pose en $\mathbf{t_{-1}}$ et est déclaré résolu en $\mathbf{t_0}$

L'application de ces tests ne nous a pas pleinement satisfaite pour caractériser les emplois métadiscursifs (reformulation, correction, résomption…) dans la mesure où ces derniers s'appliquent tous, comme leur nom l'indique, à un problème discursif, la résolution de celui-ci intervenant au moment de la réalisation de *enfin*. Ainsi, afin d'affiner davantage notre classement des emplois métadiscursifs, nous nous sommes également fondée sur l'observation de critères constructionnels tels que la présence d'un élément résomptif ou d'une justification faisant suite à *enfin*, ou encore, pour ne citer que ceux-là, la collocation de *enfin* avec *mais*.

En appliquant cette méthode, en reprenant des étiquettes pour dénommer des emplois lorsque celles-ci étaient disponibles dans la littérature ou en en créant de nouvelles lorsque cela s'est avéré nécessaire, nous avons abouti au classement des profils de *enfin* suivants :

- reformulation corrective
- correction argumentative (mettant en jeu deux ou trois mouvements discursifs)
- complétude discursive
- reformulation résomptive
- justification
- soulagement
- résignation
- irritation
- incompréhension

L'élaboration de ce classement de travail a donc été menée sans prise en considération de la dimension prosodique.

6.3 Méthode et résultats

Après avoir cherché à savoir si, pour un profil précis, les configurations prosodiques des occurrences correspondantes étaient identiques, nous avons rapidement conclu à une forte hétérogénéité dans les caractérisations prosodiques des données, ce constat remettant ainsi en cause notre hypothèse de lien entre un profil de *enfin* et une réalisation prosodique particulière.

Avant toutefois de conclure qu'une discrimination prosodique de la polysémie lexicale n'était en réalité pas de mise, nous avons opté pour une seconde méthodologie consistant à prendre cette fois comme point de départ la dimension prosodique. Nous avons alors cherché à observer si, pour un même profil, des configurations prosodiques proches étaient la manifestation de traits sémantiques identiques mais se situant à un niveau plus fin que celui

du profil. Nous avons finalement pu constater, dans un second temps, qu'il existait bel et bien des sous-emplois discriminables prosodiquement et que nous appellerons désormais des « emplois-types », présentant un degré de précision sémantique plus fin que celui du profilage. L'existence même de ces emplois-types est liée à l'expression d'un rapport thymique[43] ou attentionnel de la part du locuteur au discours, et en l'occurrence à ce que dit *enfin* et plus précisément à un profil de *enfin*. Nous entendons par là que lorsqu'un emploi de soulagement ou de résignation par exemple est réalisé, le locuteur exprime outre cela, autre chose, qui peut être, pour le soulagement son entière satisfaction ou bien le fait qu'il demeure malgré tout irrité (l'emploi-type alors obtenu est aisément glosable par « c'est pas trop tôt »), ou pour la résignation le fait qu'il se résigne en conservant ou non une certaine rancune. Ce trait interprétatif (*i.e.* ce rapport au profil) n'est perceptible qu'au travers de la prosodie, la seule dimension écrite d'un énoncé tel que « Enfin Paul arrive » ne permet pas de savoir si l'on a affaire à l'expression d'un soulagement manifeste ou bien d'un soulagement encore teinté d'une certaine irritation, même si l'on ajoute des points d'exclamation. Il est important de signaler que la discrimination prosodique n'est applicable qu'à partir du moment où le profil est déjà identifié à l'aide d'un calcul sémantico-pragmatique.

Sur ce mode consistant à alterner la prise en considération du sens et de la forme sonore, nous avons affiné en deux ou trois emplois-types chacun des profils de *enfin*, en définissant au cours de ce travail, la caractérisation prosodique de chacun de ces emplois-types. Celle-ci s'établit de deux manières, de par la configuration prosodique de *enfin* (on a généralement dans ce cas affaire à l'expression d'un rapport thymique) ou bien en fonction de la saillance prosodique de *enfin* par rapport à son contexte (il s'agit alors souvent de l'expression d'un rapport de nature attentionnelle). L'expression d'un rapport thymique et d'un rapport attentionnel ne sont pas exclusives. Voici le classement final des profils et emplois-types de *enfin* obtenu suite à ce travail et pour lequel nous avons systématiquement associé une configuration prosodique à un emploi-type :

- Les emplois de reformulation corrective :

1) La reformulation apportée est significative et marque une nette différence dans la force argumentative des arguments compris dans les séquences discursives connectées par *enfin* : Fo montante sur *enfin* + gradation possible par la présence de pauses en collocation.

2) La reformulation apportée est peu significative et marque une faible différence entre les deux arguments en question : Fo descendante sur *enfin.*

- L'emploi de signalement d'une inadéquation lexicale : Fo descendante sur *enfin* + antéposition de *enfin* par rapport à la formulation initiale.

20. Les emplois de correction argumentative :

1) Le dernier argument (introduit par *mais enfin*) est présenté comme ayant une grande force argumentative par rapport à celui qui précède.

2) Le dernier argument (introduit par *mais enfin*) est présenté comme n'ayant pas une grande force argumentative par rapport à celui qui précède.

Plus la force argumentative du dernier argument sera présentée comme forte, plus les paramètres prosodiques permettant de l'exprimer (mélodie montante sur *enfin*, mélodie montante ou en forme de cloche sur l'une des syllabes de *enfin*, présence de pause(s) en collocation avec *mais enfin,* saillance prosodique de *enfin* par rapport à son contexte) apparaîtront simultanément.

[43] Relatif à l'humeur du locuteur.

21. Les emplois de justification :

1) L'emploi de légitimation face à un désaccord implicite de l'interlocuteur : Fo montante sur *enfin*.

2) L'emploi de précision des propos tenus : Fo descendante sur *enfin*.

- Les emplois de reformulation résomptive :

1) L'emploi de synthèse pertinente : Fo montante sur *enfin*.

2) L'emploi de clôture de l'énoncé : Fo descendante sur *enfin*.

2. Les emplois de complétude discursive :

1) Mise en place d'une hiérarchisation attentionnelle des éléments : Fo montante sur *enfin* s'accompagnant d'une rupture prosodique de *enfin* par rapport à son contexte.

2) Mise en évidence de la complétude du discours : Fo montante sur *enfin* souvent accompagnée d'une post-position du connecteur.

3) Absence de hiérarchisation attentionnelle des éléments : Fo descendante sur *enfin* ne s'accompagnant pas d'une rupture prosodique de *enfin* par rapport à son contexte.

3. Les emplois de « problème résolu » (soulagement) :

1) L'emploi de soulagement manifeste : Fo montante sur *enfin* (+ cloche sur -*in* s'il y a une forme d'insistance)

2) L'emploi de soulagement masqué (avec l'expression d'une irritation résiduelle) : Fo descendante sur *enfin* (+ cloche sur -e*n* s'il y a une forme d'insistance)

3) La transition vers le soulagement : Fo montante + cloche sur -*en*.

Une saillance prosodique du connecteur par rapport à son contexte, ou la présence de pauses en collocation avec le connecteur permettent d'exprimer une gradation dans l'expression du sentiment.

- Les emplois de résignation :

1) La résignation de bonne grâce : Fo descendante sur *enfin*.

2) La résignation de mauvaise grâce : Fo montante sur *enfin*.

3) L'« emploi de transition vers l'acceptation de la situation » : Fo descendante sur *enfin* + mélodie en forme de cloche sur -*en*.

Une saillance prosodique du connecteur par rapport à son contexte, ou la présence de pauses en collocation avec le connecteur permettent d'exprimer une gradation dans l'expression du sentiment.

Les emplois de mécontentement :

1) L'agressivité, l'irritation, l'indignation, le reproche adouci de type *voyons*.

2) L'incompréhension, la surprise, l'incrédulité, l'inquiétude.

3) La lassitude, le sentiment blasé, le désespoir, la fatalité.

2. Le sentiment d'incompréhension :

Celui-ci ne constitue pas un emploi-type mais peut être un trait sémantique constitutif de différents emplois-types, relevant eux-mêmes de différentes interprétation-types. Il se manifeste par un assourdissement partiel ou total de l'occurrence selon le degré d'incompréhension exprimé.

Notre travail nous a également conduite, en mettant au jour deux niveaux de sens que sont le profil et l'emploi-type, à constater des surgénéralisations dans la manière dont nous avons dénommé certains profils. Ainsi le terme de « soulagement » convient davantage à l'emploi-type de soulagment manifeste et celui de « résignation » s'apparentant plutôt pour sa part à de la résignation de mauvaise grâce. Pour cette raison, nous avons renommé le profil de soulagement, de manière plus neutre, par le syntagme « problème résolu » et nous modifierons également dans nos travaux ultérieurs l'étiquette initialement appelée « résignation ». Nous avons par ailleurs été amenée à modifier le nom du profil « irritation » par celui de « mécontentement » après avoir pu constater qu'il était possible pour un locuteur d'exprimer sa désapprobation sans marque d'irritation. La catégorie des emplois de mécontentement s'est avérée à telle point hétérogène du point de vue de l'expression des sentiments que nous n'avons pas été en mesure d'associer des configurations prosodiques à des emplois-types. Nous avons simplement pu proposer, à ce stade de nos recherches, un classement affiné des sous-emplois relevant de cette catégorie.

Ces observations soulignent tout l'intérêt de la pratique d'une linguistique de corpus, qui permet au chercheur d'avoir accès à des emplois qu'il n'imaginait pas de prime abord. C'est notamment grâce à notre fréquentation poussée des données extraites de discours authentiques que nous avons pu isoler l'emploi, très peu fréquent, de signalement d'une inadéquation lexicale de *enfin*, se rapprochant en ce sens d'un emploi de *disons*.

Ajoutons que la langue étant par nature intrinsèquement gradable, comme l'a souligné Anscombre (1995), il est fréquent d'observer des niveaux intermédiaires - ou emplois de transition – par exemple entre une manifestation de soulagement manifeste et l'expression d'un soulagment teinté d'une forte irritation résiduelle. Si cette gradabilité inhérente participe à la richesse de l'expressivité, elle rend néanmoins plus ardue l'association d'une prosodie à un emploi-type dans la mise en place du processus de discrimination.

7 ELARGISSEMENT

Nous avons élargi notre étude aux items *quelques* et *oui* dans le but de vérifier si l'expression par le locuteur d'un rapport au profil se manifestait de la même manière sur d'autres types d'unités. Nous allons présenter brièvement nos résultats ici.

7.1 *Quelques*

L'adjectif *quelques* peut notamment présenter :

- une lecture minorante comme dans « j'ai seulement *quelques* minutes à te consacrer » où la quantité est présentée comme faible.
- une lecture majorante comme dans « il y a *quelques* objets très intéressants » où la quantité est présentée comme significative.

L'étude prosodique de cette unité a permis d'observer que, selon la saillance prosodique de *quelques* par rapport à son contexte, il est possible, non pas de discriminer les deux types d'emplois mentionnés, mais de préciser pour chacun d'entre eux si la quantité exprimée mérite ou non de l'attention, c'est-à-dire que le locuteur a la possibilité de signaler si la quantité en question a de l'importance dans l'argumentation.

7.2 *Oui*

L'analyse de l'adverbe *oui* a pour sa part permis de mettre au jour le fait que le locuteur pouvait, par la réalisation prosodique de *oui*, exprimer différents degrés d'acceptation pouvant aller de l'accord plein et sincère à l'expression d'une forte réticence, marquée par une cloche mélodique sur *oui*.

8 GÉNÉRALISATION ET FORMAT DE REPRÉSENTATION

L'analyse de *enfin* ainsi que les études complémentaires que nous avons menées (voir Petit, 2009) nous amènent à penser qu'il existe une contrainte générale qui pèse sur le discours oralisé : il existe deux niveaux systématiquement exprimés et qui sont sources de polysémie :

13. ce que l'on dit (qui correspond aux profils)
14. ce que l'on en dit (qui correspond à un rapport ou à un « commentaire » sur ce qui est dit, commentaire marqué prosodiquement)

Nous pensons en effet que dès lors qu'un locuteur utilise le discours oral, il ne peut pas ne pas exprimer en même temps un rapport à ce qu'il dit, le rapport en question pouvant exprimer la neutralité. Nous ne sommes pas, à ce stade de notre recherche, en mesure de déterminer la part du caractère volontaire ou trahi de la manifestation de ce rapport.

Sur la base de ces observations, nous proposons le format de représentation des sens lexicaux (ou emplois-types) suivants :

Morphème (qui code la signification) :
Profil 1 (résultant d'un calcul sémantico-pragmatique) :

- Emploi-type 1 (résultant de l'expression d'un rapport au profil)
- Emploi-type 2
- Emploi-type 3

Profil 2

- Emploi-type 1
- Emploi-type 2…

Profil 3

- Emploi-type 1…

Nous pouvons l'illustrer de la manière suivante à l'aide du profil de « problème résolu » de *enfin* :

Enfin (« Indication qu'un problème se pose à un moment t et est résolu en t+1 ») :
Problème résolu (le problème situationnel est résolu au moment où le locuteur réalise *enfin*) :

- Soulagement manifeste (le locuteur est content que le problème soit résolu)
- Soulagement masqué accompagné d'une irritation résiduelle (le locuteur est encore irrité qu'un problème se soit posé, même s'il est résolu au moment où il parle)

Il est possible de faire figurer à l'aide de ce format de représentation tous les emplois-types de *enfin* que nous avons présentés dans notre classement final ainsi que les emplois-types de *quelques* et de *oui* à partir du moment où le rapport au profil exprimé est identifié. Plus globalement, partant du principe que l'expression d'un rapport est générale en discours, thèse

qui mérite d'être étayée par une grande diversité d'analyses portant sur des types d'objets différents et fondée sur des corpus de travail importants, nous estimons que ce format pourrait s'appliquer à un grand nombre d'unités lexicales du français. Ce type de représentation offre en outre la possibilité de faire figurer une cohérence sémantique en représentant explicitement les liens entretenus entre les emplois-types.

Ce travail nous a également naturellement conduite à nous interroger sur la notion de signifiant. Le phénomène de polysémie peut être décrit comme l'association de plusieurs signifiés à un même signifiant. Or, la mise au jour de l'existence d'emplois-types, prosodiquement discriminables, nécessite de distinguer la forme phonématique de *enfin* (commune à tous les profils et que l'on représente à l'aide de l'API) de ce que nous appellerons le « signifiant phonologique » qui associe la forme phonématique de *enfin* à une configuration prosodique particulière. Si l'on considère le signifiant phonologique, alors la notion de polysémie telle que décrite ici disparaît au niveau des emplois-types puisque des signifiants phonologiques différents sont associés à des signifiés liés mais distincts (soulagement manifeste, soulagement masqué). Rappelons toutefois qu'il est nécessaire d'avoir au préalable identifié le profil de *enfin* par un calcul sémantico-pragmatique avant de pouvoir procéder à toute discrimination prosodique du lexique. Voici à présent une représentation formelle de nos propos à partir de l'emploi-type de soulagement manifeste :

Soulagement manifeste :
Signifiant phonologique :
1.1.1. Forme phonématique : *enfin* (API)
1.1.2. Forme prosodique : Fo montante sur *enfin* (pouvant s'accompagner d'une cloche mélodique sur *–fin*)
Interprétation :
ii. Profil : un problème s'est posé, il est résolu au moment de l'énonciation de *enfin*
iii. Rapport exprimé : satisfaction que le problème soit résolu
Statut constructionnel : modifieur de prédicat

Au cours de nos analyses, nous n'avons pas observé de corrélation entre la syntaxe et la prosodie de *enfin*. La caractérisation grammaticale de cette unité est par ailleurs très complexe notamment en raison de la forte hétérogénéité contextuelle qu'elle est susceptible de présenter. Il est en effet tout à fait possible de rencontrer par exemple un emploi-type de soulagement masqué dans des contextes très variés (sans énoncé avant et/ou après etc.) mais également de rencontrer des emplois-types de soulagement manifeste et de soulagement masqué dans des contextes identiques.

9 PERSPECTIVES

Cette dichotomie entre le signifiant phonématique et le signifiant phonologique signalée, nous soumettons l'hypothèse que le repérage automatique des différents sens (nous considérons l'emploi-type comme le sens et le profil comme un accès certes indispensable mais intermédiaire au sens) d'une même unité s'en trouverait simplifié dès lors que la configuration prosodique est prise en considération au niveau de l'unité lexicale. Cette possibilité s'avère d'autant plus intéressante qu'il s'agit de traiter d'unités-clés dans la mise en place d'une argumentation et dans l'organisation discursive : les connecteurs.

L'application du processus de discrimination prosodique que nous avons présenté permet par ailleurs d'atteindre un degré de précision sémantique important, notamment, à l'instar de

Cadiot & Visetti (2001), par l'adoption d'un modèle ternaire (signification, profil, emploi-type) et non plus binaire (signification, sens) de représentation du sens.

Etant donnée la forte corrélation entre l'expression d'un rapport au profil et la dimension argumentative de la langue, il est également possible d'avoir accès, par le biais de la prosodie, aux intentions des locuteurs.

10 RÉFÉRENCES

Anscombre J. C. (éd.) (1995). *Théorie des topoï.* Paris : Kimé.

Barnes B. (1995). « Discourse Particles in French Conversation: *(eh) ben, bon* and *enfin* ». *The French Review n° 68*, 5., p. 813-821.

Beeching K. (2007). « La co-variation des marqueurs discursifs *bon*, *c'est-à-dire, enfin, hein, quand même, quoi* et *si vous voulez* : une question d'identité ? ». *Langue française n° 154*.2, p. 78-93.

Beeching K. (2002). *Gender, politeness and pragmatic particles in French.* Amsterdam et Philadelphia : John Benjamins.

Beeching K. (2001). « Repair strategies and social interaction in spontaneous spoken French: the pragmatic particle *enfin* ». *Journal of French Language Studies n° 11*, 1., p. 23-40.

Beeching K. (2000). « La fonction de la particule pragmatique *enfin* dans le discours des hommes et des femmes ». Dans N. R. Armstrong, C. Bauvois et K. Beeching (éds.), *Femmes et français*. Paris : L'Harmattan.

Bertrand R. et Chanet C. (2005). « Fonctions pragmatiques et prosodie de *enfin* en français spontané *». Revue de Sémantique et de Pragmatique n° 17*, p. 41-68.

Buchi E. et Städtler T. (2008). « La pragmaticalisation de l'adverbe *enfin* du point de vue des romanistes (« Enfin, de celui des francisants qui conçoivent leur recherche dans le cadre de la linguistique romane ») ». Dans J. Durand, B. Habert et B. Laks (éds.), *Congrès mondial de linguistique française (Paris, 9-12 juillet 2008). Recueil des résumés, CD-ROM des actes.* Paris : Institut de linguistique française. p. 159-171.

Cadiot A. et al. (1985). « *Enfin*, marqueur métalinguistique ». *Journal of pragmatics n° 9*, p. 199-239.

Cadiot P. et Visetti Y. M. (2001). « Motifs, profils, thèmes : une approche globale de la polysémie ». *Cahiers de lexicologie n° 79*, 2001-2, p. 5-46.

Chanet C. (2003). « La forme « enfin » en français parlé contemporain : vers une typologie des statuts et des emplois ». Dans *VIII Simposio de Comunicación social (Santiago de Cuba, 20-24 janvier 2003). Actas I.* Santiago de Cuba : Centro de Lingüística Aplicada. p. 394-399.

Cloiseau G. (2007). *Une redéfinition de la métaphoricité à l'oral : mise en place d'outils d'analyse par une approche de corpus contrastive.* Thèse de Doctorat, Université d'Orléans.

Dostie G. (2004). *Pragmaticalisation et marqueurs discursifs, analyse sémantique et traitement lexicographique.* Bruxelles : Duculot.

Fónagy I. (1983). *La vive voix: essai de psychophonétique.* Paris : Payot.

Franckel J.J. (1987). « *Fin* en perspective : *finalement, enfin, à la fin* ». *Cahiers de Linguistique Française n° 8*, p. 43-68.

Fuchs C. (1996). *Les ambiguïtés du français.* Paris/Gap : Ophrys.

Hansen M-B Mosegaard (2005a). « From prepositional phrase to hesitation marker: The semantic and pragmatic evolution of French *enfin* ». *Journal of Historical Pragmatics*, 6, 1, p. 37-68.

Hansen M-B Mosegaard (2005b). « A comparative study of the semantics and pragmatics of e*nfin* and *finalement*, in synchrony and diachrony ». *Journal of French Language Studies*, 15, 2, p. 153-171.

Hwang Y. A. (1993). « *Eh bien, alors, enfin* et *disons* en français parlé contemporain ». *L'Information Grammaticale n° 57*, p. 46-48.

Luscher J. M. et Moeschler J. (1990). « Approches dérivationnelles et procédurales des opérateurs et connecteurs temporels : les exemples de *et* et de *enfin ». Cahiers de Linguistique Française 11*, p. 77-104.

Nemo F. (2000). « *Enfin*, *encore*, *toujours* entre indexicalité et emplois ». Dans Englebert A. et al. (éds.), *Actes du XXIIe Congrès international de linguistique et de philologie romanes*, (Bruxelles, juillet 1998). Tübingen : Max Niemeyer Verlag, vol. 7. p. 499-511.

Noda H. (2005). « L'emploi des mots du discours et la prosodie: le cas de hein ». *Proc. Interface Discours Prosodie 2005.* Aix-en-Provence.

Paillard D. (2003). « À propos de *enfin* ». Dans B. Combettes, C. Schnedecker et A. Theissen (éds.), *Ordre et distinction dans la langue et le discours. Actes du Colloque international de Metz (18, 19, 20 mars 1999).* Paris : Champion. p. 387-408.

Petit M. (2009). *Discrimination prosodique et représentation du lexique : application aux emplois des connecteurs discursifs.* Thèse de Doctorat, Université d'Orléans. 498 p.

Razgouliaeva A. (2004). « Les combinaisons des connecteurs *mais enfin* et *mais de toute façon* ». Dans C. Rossari (éd.), *Autour des connecteurs. Réflexions sur l'énonciation et la portée.* Berne : Peter Lang. p. 157-180.

Razgouliaeva A. (2002). « Combinaison des connecteurs *mais* et *enfin* ». *Cahiers de linguistique française n° 24,* p. 143-168.

Rossari C. et al. (2004). *Autour des connecteurs. Réflexions sur l'énonciation et la portée.* Berne : Lang.

Rossari C. (2000). *Connecteurs et relations de discours : des liens entre cognition et signification.* Nancy : Presses Universitaires de Nancy.

Rossari C. (1997). *Les opérations de reformulation. Analyse du processus et des marqueurs dans une perspective français-italien.* Berne : Lang.

Vincent D. et Demers M. (1994). « Les problèmes d'arrimage entre les études discursives et prosodiques. Le cas du « là » ponctuant ». *Langues et linguistique*, 20, p. 201-212.

Waltereit R. (2007). « A propos de la genèse diachronique des combinaisons de marqueurs. L'exemple de bon ben et enfin bref ». *Langue française n° 154.2,* p. 94-109.

ÉTUDE ET TRAITEMENT AUTOMATIQUE DE L'ANGLAIS DU XVII^E SIÈCLE : OUTILS MORPHOSYNTAXIQUES ET DICTIONNAIRES

Hélène Pignot et Odile Piton
Laboratoire SAMM-Marin Mersenne – Université Paris1 Panthéon-Sorbonne

RÉSUMÉ

Après avoir exposé la constitution du corpus, nous recensons les principales différences ou particularités linguistiques de la langue anglaise du XVII^e siècle, les analysons du point de vue morphologique et syntaxique et proposons des équivalents en anglais contemporain (AC). Nous montrons comment nous pouvons effectuer une transcription automatique de textes anglais du XVII^e siècle en anglais moderne, en combinant l'utilisation de dictionnaires électroniques avec des règles de transcriptions implémentées sous forme de transducteurs.

ABSTRACT

In this article, we record the main linguistic differences or singularities of 17th century English, analyse them morphologically and syntactically and propose equivalent forms in contemporary English. We show how 17th century texts may be transcribed into modern English, combining the use of electronic dictionaries with rules of transcription implemented as transducers.

1 INTRODUCTION

L'anglais du XVII^e siècle présente de nombreuses particularités orthographiques, syntaxiques et lexicales qui en font tout le charme mais également la difficulté. Même pour un locuteur natif, la lecture de ces beaux textes qui font partie du patrimoine littéraire et historique de l'humanité, n'est pas chose aisée. Les ouvrages du XVII^e siècle réédités pour les étudiants ou pour le grand public sont donc assortis d'un riche apparat critique, et de multiples notes concernant le lexique et éclairant les circonstances historiques de leur rédaction.

Notre domaine de recherche conjugue deux passions, pour l'histoire anglaise et européenne au XVII^e siècle et pour le récit de voyage. Six années durant nous avons pu redécouvrir les érudits et voyageurs français et anglais en Grèce et en Anatolie (Pignot 2007, 2009). En traduisant les textes anglais pour l'édition française ou en les établissant pour l'édition anglaise, nous avons pu mesurer leur difficulté et leur richesse. Une question a jailli dans notre esprit : comment rendre ces textes plus accessibles aux lecteurs modernes non spécialistes et aux étudiants, comment éviter qu'ils ne soient rebutés par une langue qu'ils jugeront peut-être archaïque et difficile à comprendre ?

Nous avons donc souhaité œuvrer au développement d'un outil automatique permettant d'appréhender la langue de cette époque. A partir de notre corpus nous recensons les différences ou particularités linguistiques de la langue de cette époque, les analysons du point de vue morphologique et syntaxique et proposons des équivalents en anglais contemporain (AC). Notre collaboration, qui a débuté en 2006, visait d'abord à définir des outils permettant

de formaliser les données linguistiques nécessaires à la traduction diachronique automatique (dictionnaires et grammaires bilingues). Depuis, nous avons construit, appliqué, testé et affiné ces outils, et présentons ici nos résultats.

Le présent article comportera trois volets, l'un présentant le corpus, l'autre consacré à l'analyse morphologique et syntaxique automatique de l'anglais du XVII^e siècle, et le dernier à l'étude du lexique, en particulier à la constitution d'un dictionnaire électronique recensant les particularités lexicales de l'anglais du XVII^e siècle (mots archaïques ou dont le sens a évolué en anglais moderne) et proposant – quand cela est possible – un équivalent en AC. Notre objectif ultime est de créer un dictionnaire de l'anglais du XVII^e siècle, ainsi que des outils facilitant la lecture des récits de voyage du XVII^e siècle quelle que soit leur aire géographique, et leur accès, par exemple pour le grand public ou pour un public d'étudiants (les implications pédagogiques seront évoquées dans la conclusion).

2 CONSTITUTION ET EXPLOITATION DU CORPUS

Pour les chercheurs, de nombreux textes du du XVII^e siècle sont accessibles grâce à la base de données EEBO (*Early English Books Online*). Les textes sont ainsi téléchargeables au format 'pdf' (image). Au moment où nous avons constitué le corpus, nous n'avions pas accès à EEBO ni à des logiciels d'OCR. À l'époque de la sélection de ces textes nous avons fait le choix de retaper tous les extraits relatifs à la Grèce.

Constatant une évolution des fontes d'imprimerie, nous avons dû opérer quelques modifications. Au cours de notre opération de saisie nous avons remplacé le « s long » par des s et remplacé l'éperluette par « and » (on ne peut donc rechercher les occurrences de ces caractères dans notre corpus). Il faut noter que la différentiation entre i et j, u et v n'est effective qu'à partir de 1634, comme le remarque Robert Lass (Hogg, Lass et al. 1992). On recommande aux imprimeurs d'éviter certains doublements de consonnes et l'ajout de e muet final mais les pratiques varient, voir les **Images1, 2 et 3**.

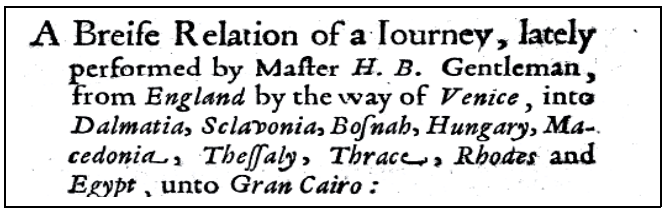

A Breife Relation of a Iourney, lately performed by Maſter *H. B.* Gentleman, from *England* by the way of *Venice*, into *Dalmatia, Sclavonia, Boſnah, Hungary, Macedonia, Theſſaly, Thrace, Rhodes* and *Egypt*, unto *Gran Cairo:*

Image 1 : Exemple de typographie de l'anglais du XVII^e

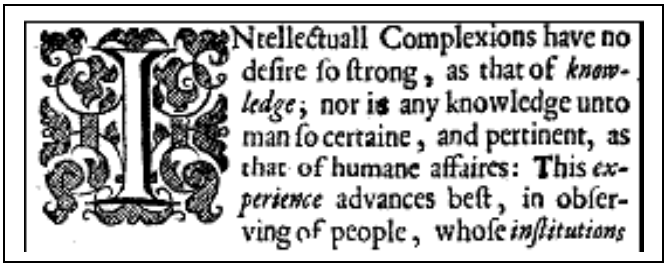

INtellectuall Complexions have no deſire ſo ſtrong, as that of *knowledge*; nor is any knowledge unto man ſo certaine, and pertinent, as that of humane affaires: This *experience* advances beſt, in obſerving of people, whoſe *inſtitutions*

Image 2 : Exemple de typographie de l'anglais du XVII^e

XVII[e]	AC	XVII[e]	AC	XVII[e]	AC
&c	etc.	ſeuerall	several	ſoueraignтie	sovereignty
vnder	under	*Moſes.*	Moses	ſeaſon	season
Iewes	Jews	haue	have	ſlauery	slavery
approoued	approved	ioyneth	joins	diſpoſſeſſeth	dispossesses

Image 3 : Typographie de l'anglais du XVII[e]

Les auteurs qui constituent notre corpus sont George Sandys, Henry Blount, John Ray, Paul Rycaut, Thomas Smith et George Wheler. Chronologiquement le premier est le poète George Sandys (1578-1644), fils d'un ecclésiastique, Edwin Sandys, qui fut archevêque d'York. C'est peut-être pour se consoler d'un mariage malheureux et de ses diverses complications que Sandys décida de s'embarquer pour le Levant. Selon l'historien Hugh Trevor Roper, George Sandys désirait l'union des églises et portait de ce fait un certain intérêt à la situation de l'Église grecque dans l'Empire ottoman. L'itinéraire emprunté par notre voyageur lui permet de visiter la France, le Nord de l'Italie, la Turquie (qu'il explore une année durant), l'Égypte, la Palestine, Chypre, la Sicile, Naples, et Rome. Sa relation de voyage, qui parut sous le titre, *A Voyage to the Levant*, fit florès, connaissant sept rééditions au cours du XVII[e] siècle. Il y consacre quelques pages aux coutumes des Grecs, à leur foi, à leur mode de vie, à leurs coutumes, et à leur langue.

Le second auteur est Henry Blount (1602-1682), avocat de profession, qui s'embarqua à Venise sur un navire en direction du Levant, en compagnie de voyageurs turcs et juifs. Désireux de découvrir la Turquie et de comprendre les raisons de sa puissance, il en profita pour faire connaissance des peuples qui vivent sous le joug ottoman, « les Grecs, les Arméniens, les Francs, les Tsiganes, et particulièrement les Juifs ». Sa relation de voyage, *A Voyage into the Levant*, fut publiée pour la première fois en 1636, et connut huit éditions entre 1636 et 1671. Ce qui retient l'attention de ce voyageur plus sceptique de nature, ce sont surtout les renégats du christianisme.

John Ray (1627-1705) était pour sa part botaniste. Il n'a pas visité l'Empire ottoman mais a parcouru l'Europe trois années durant, découvrant la Hollande, l'Allemagne, la Suisse, l'Italie, la Sicile et Malte. Dans son ouvrage *A Collection of Curious Travels and Voyages through the Levant* (1693), Ray réunit les contributions de divers voyageurs dans l'Empire ottoman et résume les chapitres que Pierre Belon, botaniste français et voyageur au Levant au milieu de XVI[e] siècle consacre au Mont Athos dans un de ses récits de voyage.

Quant à Paul Rycaut (1629-1700), il fit carrière dans la diplomatie d'abord comme secrétaire de l'ambassadeur du roi d'Angleterre Charles II à Constantinople auprès du Sultan Mahomet IV, puis de 1667 à 1678, comme consul d'Angleterre à Smyrne. Cette position d'observateur privilégié lui permettra de publier en 1679 *The Present State of the Greek et Armenian Churches*, qui n'est pas seulement un témoignage sur ces deux Églises et sur leurs coutumes, mais également une relation de voyage puisque le texte comprend la description de certains lieux visités par l'auteur tels l'Asie Mineure et le Mont Athos. Rycaut souhaitait également l'union des Églises et rédigea ce livre sur les instances du roi Charles II, qui désirait parfaire sa connaissance de la foi des orthodoxes grecs.

Thomas Smith (1638-1710) étudia les langues orientales et fut *fellow* de Madgalene College à Oxford. En qualité de chapelain, il accompagna Sir Daniel Harvey, ambassadeur d'Angleterre à Constantinople et demeura trois années à son service (de 1668 à 1671). En 1672 il visita les sept Églises d'Asie et en donna une description dans un ouvrage rédigé en latin et publié en 1694. Dans sa correspondance, lui aussi s'interroge sur la possibilité d'une

union entre l'Église d'Angleterre et celle de Grèce, et il publie en 1680 un ouvrage intitulé *An Account of the Greek Church*, où il expose les rites et les doctrines de cette Église.

Pour clore cette liste, le botaniste anglais George Wheler (1650-1723), partit à la découverte de la Grèce en compagnie du médecin et archéologue lyonnais Jacob Spon (1647-1685). Ils publièrent chacun un récit de leur voyage (Wheler, 1682). À son retour de Grèce, Wheler fut ordonné pasteur. Il fut l'un des premiers membres des *Society for the Propagation of the Gospel* et *Society for Promoting Christian Knowledge*, et la Grèce lui semblera un terrain propice à l'activité missionnaire anglicane.

Ainsi, ce qui rapproche nos auteurs, qui voyagent à des moments historiques très différents, c'est la curiosité pour la Grèce et le désir de comprendre ce qui différencie les Grecs des autres chrétiens. Ces différences seraient-elles de nature à empêcher une union avec l'Église d'Angleterre ? Rycaut et Smith se posent clairement la question. Toutes ces précisions historiques sont indispensables pour nous permettre de comprendre pourquoi un certain nombre des exemples de notre troisième partie sera emprunté au vocabulaire religieux : la société grecque dépeinte par nos voyageurs est organisée autour de ses croyances religieuses, la vie est rythmée par le calendrier ecclésiastique, les fêtes religieuses, les carêmes, les prières et tous les rites de l'Église.

Ces textes nous fournissent un échantillon linguistique qui balaie tout le siècle (de 1615 à 1693). Ce corpus est multilingue. Il comporte de nombreuses citations en grec ancien, en latin, des emprunts plus ou moins nombreux selon les langues, au turc, au grec, à l'italien et à l'espagnol. Il nous permet de dresser un inventaire des différences orthographiques, lexicales et syntaxiques entre l'anglais du XVII^e^ siècle et l'anglais moderne ; on remarquera une certaine cohérence.

3 ÉTUDE DIACHRONIQUE : MODÉLISATION DES ÉVOLUTIONS DE LA LANGUE

Toponymes et ethnonymes étaient souvent différents en anglais du XVII^e^ siècle, ainsi que leur orthographe, ce qui engendre un double dépaysement pour le lecteur, plongé dans un lointain univers culturel, d'ailleurs très souvent qualifié d'« oriental » par nos voyageurs lors même qu'il s'agit des Grecs, et doté d'une toponymie différente et pour le moins déroutante parfois ! Donnons quelques exemples : Nice est un autre nom de Nicée en Asie Mineure, Romagnia désigne la Grèce et les Balkans, le Maina est le sud du Péloponnèse (Mani en anglais moderne), la Croatie s'appelle Sclavonia, les îles de la mer Egée situées entre la Grèce et la Turquie ou îles de l'Archipel se nomment aussi « the Arches ».

Le vocable « Hollanders » désigne les Hollandais, « Grecians » les Grecs (et pas seulement ceux de l'Antiquité comme en anglais moderne), les Zynganaes ou « Zinganies », les tziganes.

Parmi les termes historiquement datés, nous avons également fait le relevé de toutes les unités de poids et mesure, de distance spécifiques au XVII^e^ siècle ainsi que des monnaies, qui ne sont guère parlantes pour un lecteur non averti. Nous avons aussi répertorié les titres étrangers. Voir la Table1.

Catégorie	Exemple	Définition
Monnaie	*A zecchine* *A hungar* *A sultany* *An asper* *A ducat*	Sequin, pièce vénitienne Pièce hongroise Pièce turque en or Pièce turque en argent Pièce italienne en or ou en argent utilisée dans toute l'Europe
Unité de poids	*An oque*	1kg 250
Unité de distance	*A league* *A furlong*	circa 4 km 200 m
Titre ou profession turc	*Bassa* *A cadi* *A dragoman* *Keslar-Agafi*	le Pacha un juge un interprète Eunuque noir qui surveille les femmes du Sérail
Titre grec	*Egoumen*	le Père Abbé d'un monastère
Titre italien	*the Grand Signior* *the Bailo*	le Sultan le Baile vénitien

Table 1 : Termes historiques: exemples

Nous avons fait le recensement de tous les mots étrangers et proposé un équivalent en anglais moderne. Les emprunts ont une fonction d'authentification du témoignage et manifestent le degré d'implication du voyageur dans la société où il vit. En recourant à ces mots empruntés au grec et au turc, nos auteurs veulent montrer la spécificité des us et des coutumes décrits (il y a de très nombreux termes grecs et citations en grec dans les textes de Rycaut et de Smith en particulier). La Table 2 récapitule quelques-uns de ces emprunts.

Langue	Emprunts lexicaux	Anglais contemporain
Arabe	*salam'd;'salamed'*	saluted
Français	*randevouzes*	pluriel du mot français rendez-vous, utilisé pour le rendez-vous galant en AC

Grec	*antidoron*	blessed bread
	comparos	godfather
	diataxis	order
	douleia	veneration of a saint or of relics and icons
	Eikonomachoi, or Eikonoklastai	those who oppose the worship of images or destroy them
	kaloir (kalogieros)	a good elder: a Greek Orthodox monk
	kosmokratores	Emperors of the world
	latinophrones	latinizing
	latreia	worship of God
	mamoukode	a ghost
	metanoïa	repentance
	metousiosis	modification
	Paranomoi	flagitious persons, and transgressors of the laws and canons of the Church
	somatikos	corporal
	sponsalia	the betrothal ring
	ta trimera	third day after death, on which prayers are said for the departed soul
	tèn diairesin apo tês alethe	a disunion from the truth
	tó déma	"tying up a man from accompanying with any woman" a spell to make a man impotent
	to katasphragisai to paidion	the sealing of infants
	vroukolakas	an evil spirit
Italien	*canaglia*	a rascal
	a capriccio of the Grand Signior	a caprice of the Sultan
	Madonna di Constantinopoli	the blessed Virgin of Constantinople (an icon believed to have been painted by St. Luke)
Latin	*dissentaneous*	contrary (latin *dissentaneus)*
	flagitious	guilty of terrible crimes (latin *flagitiosus,* OED[44] 1550)
	supposititious	spurious (latin *suppositititius,* OED, 1611)
	margaritae	pearls or particles of the Holy Communion
Turc	*alempena*	Constantinople or the refuge of the world
	baratz	Commission from the Grand Signor
	bezesten	market-place
	gazi	conqueror
	harach	poll-money (tribut)
	kabin	cohabitation as opposed to marriage
	kara congia	a demon appearing in the shape of a black old man

Table 2 : Exemples de mots empruntés

Enfin la langue du XVII[e] siècle comporte un certain nombre d'archaïsmes, de formes anglicisées de mots latins et grecs totalement inusités en anglais moderne, et de mots dont le sens a évolué. La Table 3. fournit quelques exemples significatifs, les catégories sont codées N pour les noms, A pour les adjectifs, et V pour les verbes. En faire le relevé et proposer un

[44] Oxford English Dictionary

équivalent en anglais moderne facilite la compréhension de ces textes non seulement par des lecteurs dont la langue maternelle n'est pas l'anglais mais aussi par des Anglophones. Toutes ces particularités sont répertoriées dans notre dictionnaire.

Anglais du XVIIe siècle	Catégorie	Anglais contemporain
Alcoran	N	the Koran
arbitrement	N	Arbitration
chane (Arabic, khan)	N	an inn
constitute	V	to set up in an office or position of authority
declension	N	a decline
dissentaneous	A	contrary to
drubbing	N	Beating
ethnick (E)	A	pagan, heathen (grec, ethnos, OED 1470)
gossip	N	a godfather or godmother
grogoran	N	grogram, a coarse fabric of silk, mohair and wool
penitentiary	N	a spiritual father ou a penitent (selon le contexte)
pix	N	a vessel in which the consecrated bread of the Sacrament is kept
prejudicacy	N	prejudice
runnugate	N	a renegade
shash	N	a sash, a scarf worn around the waist
symbolize with someone	V	to resemble, to partake of the nature of
tenent	N	a tenet
turcism	N	Islam
upstart (E)	A	lately come into existence

Table 3 : Archaïsmes ou termes dont le sens a évolué (E)

4 MÉTHODOLOGIE : TRAITEMENT MORPHOLOGIQUE ET SYNTAXIQUE

Nous avons opté pour l'utilisation de la plate-forme linguistique NooJ. Rappelons brièvement qu'elle permet de construire des outils sous forme de dictionnaires de formes fléchies – ces formes étant toutes associées au lemme correspondant (verrai, voyais, vis sont associées au lemme voir) – ou sous forme de grammaires exprimées par des règles morphologiques ou syntaxiques. Elle effectue à la demande l'analyse lexicale d'un texte par des outils sélectionnés (dictionnaires et grammaires), ce qui l'indexe, produit en retour des informations sur le texte, notamment la liste des formes reconnues, et celle des formes non reconnues. De nombreux travaux ayant déjà porté sur la langue anglaise contemporaine, des dictionnaires et des graphes sont mis à disposition (Silberztein 1993, 2003). *A contrario*, les particularités de la langue du XVIIe ne sont pas prises en charge. C'est soit au moyen de dictionnaires, soit au moyen de grammaires que nous allons définir les outils qui vont permettre de pré-analyser automatiquement un texte du XVIIe, ainsi que les propositions de réécriture de mots ou fragments de textes en AC. Nous obtenons ainsi des entrées lemmatisées qui, après validation, vont enrichir notre dictionnaire. Notre but est double : constituer un dictionnaire et affiner nos outils de reconnaissance et de traitement.

Les entrées d'un dictionnaire électronique sont de la forme : lemme,code catégorie+trait1+ …+traitn. Les codes de catégories sont respectivement N, A, V, PRO… pour les noms, les adjectifs, les verbes et les pronoms…. Les traits syntaxiques sont insérés lors de la

compilation du dictionnaire, selon le modèle indiqué sous la forme FLX=codeflexion. Le trait +EN=xxx permet d'indiquer la transcription en AC du mot concerné. Les différents codes sont indiqués dans la documentation NooJ, qui est accessible en ligne à l'adresse suivante <www.nooj4nlp.fr>. Nous verrons que nous pouvons être conduits à établir des ordres de priorité entre certains mots, ce qui revient à constituer *plusieurs dictionnaires* et à les hiérarchiser selon un système de priorité.

Insistons sur le fait que l'approche du traitement informatique est bien différente de l'approche linguistique, en conséquence des traitements automatiques similaires peuvent concerner des aspects linguistiques très différents, et inversement des variations d'un même élément linguistique peuvent nécessiter des outils de traitement de nature totalement différente. Nous présentons une étude détaillée, et en indiquons le mode opératoire : création de nouvelles formes lemmatisées, et règles de réécriture des mots isolés et des syntagmes identifiés, que nous pouvons automatiser. Enfin nous proposerons un bilan de notre travail et des difficultés rencontrées.

Il importe de faire clairement la distinction opérée par NooJ entre grammaire morphologique et grammaire syntaxique : une *grammaire morphologique*, permet de traiter un lexème unique, tandis qu'une *grammaire syntaxique* traite de formes disjointes, lesquelles peuvent concerner des lexèmes distincts, ou appartenir au même lexème comme lors de l'utilisation de /'d/ pour remplacer une flexion ed (*dry'd* pour dried). Nous les présenterons sous forme de graphes.

L'observation des transformations, et leur aspect rare ou systématique donne lieu à deux traitements qui sont complémentaires : une modification isolée est traitée par une entrée dans le dictionnaire électronique, qui comportera la réécriture du mot en AC, tandis qu'une modification qui nous paraît avoir un caractère aspect répétitif, est écrite sous forme de règle permettant d'identifier des lexèmes dans le texte, et de produire automatiquement des propositions d'entrées pour le dictionnaire. Nous verrons l'importance de l'ordre d'application des règles ainsi décrites.

L'examen de l'anglais du XVII^{e} nous permet d'observer des différences orthographiques telles que le redoublement de consonnes, des différences de préfixes et de suffixes, l'ajout ou la suppression de e muets. Certaines lettres ont deux formes : u et w, i et j comme en latin classique. Certains mots sont dissociés en deux lexèmes alors qu'en anglais moderne ils sont concaténés, comme "for ever" ou "sun-set", ou inversement comme dans le mot monyworth, ancienne forme de l'expression "money's worth". La ponctuation est différente : les deux-points sont utilisés soit pour marquer la fin d'une phrase, soit pour marquer une pause dans la phrase. L'éperluette (&) remplace souvent la conjonction "et".

Nous présentons en premier lieu le traitement de formes grammaticales, puis nous traiterons des variations morphologiques et des expressions disjointes.

4.1 Les modifications grammaticales

–Les noms et pronoms : nombre de ces mots comportent deux composants séparés, parfois reliés par un trait d'union, alors qu'en AC ils sont concaténés (par exemple any thing, any one, church-yard, ou Arch-Angel) ou au contraire, North wind est écrit Northwind. La transformation de deux lexèmes reliés par un trait d'union en lexème unique se fait par un simple graphe qui concatène les deux éléments. La Figure 2 présente le graphe qui permet de transcrire automatiquement mid-land en <midland> ; countrey-men en <countrymen>.

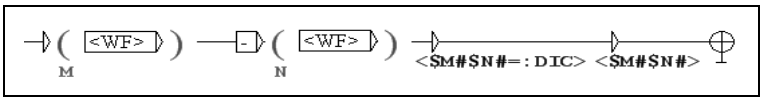


Figure 2 : Suppression du trait d'union

Inversement la transformation d'un lexème unique en mot composé ou en expression disjointe se fait au moyen d'entrées dans le dictionnaire. Soient monyworth,N+EN="money's worth" et Northwind,N+s+EN="North wind".

–Le génitif, qui s'écrivait sans apostrophe : "from the *womens* apartment"; "out of their wives and childrens mouths" commence à s'orthographier /'s/; mais outre le génitif —dans "the grand Signior's women"—, on rencontre la flexion /'s/: pour marquer le pluriel interlingual de noms étrangers, comme dans *Egoumeno*'s, *Bassa*'s (Pashas), ou *piazza*'s qui sont reconnaissables par le graphe présenté en Figure 3.

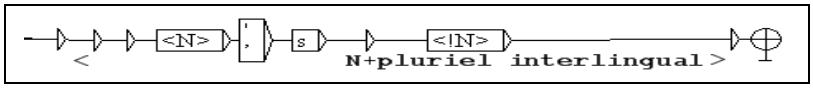


Figure 3 : Pluriel interlingual

Le premier graphe présenté en Figure 4 est un graphe morphologique qui permet de marquer un mot inconnu terminé par un s, comme anothers (dans "one anothers company"), ou childrens (dans "their wives and childrens mouths"). Remarquons que wives est également un génitif pluriel, il est analysé comme un pluriel et la conjonction and permettra sa reconnaissance par le graphe syntaxique. Celui-ci identifie la succession <N+gen_sax> <N>, ou <PRO+gen_sax> <N>, soit childrens mouths et anothers company. Il propose les transcriptions : <another's company> et < wives's and children's mouths>[45].

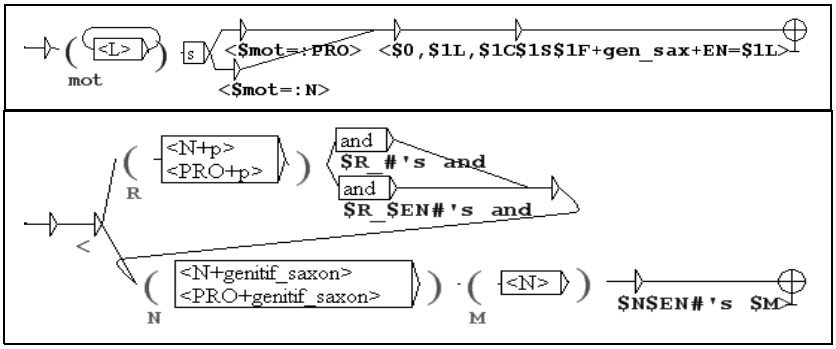


Figure 4 : Identification du génitif saxon et insertion de /'s/

Notons que mans dans « no other mans errors could draw » n'est pas identifié comme un génitif, car il est faussement reconnu comme le verbe man.

– Pronoms et adjectifs.

–Les pronoms personnels peuvent s'orthographier hee, shee, wee, ye, thay au lieu de he, she, we, you, et they. Ceci se traite simplement par le dictionnaire : hee,PRO+m+s+EN=he ; shee,PRO+m+s+EN=she etc. La 2e personne du singulier –thou, thee, thine = tu, toi, tien– est

[45] wives' (GB) ou (wives's) US. Notre objectif est de repérer le génitif, et pour en rendre possible le traitement automatique, nous appliquons la règle de *The Chicago Manual of Style* qui recommande l'addition du s, ayant bien conscience qu'il s'agit d'une norme américaine et non britannique.

souvent utilisée (particulièrement dans les prières) : “We offer to thee thine of thine own[46], in all things and through all things” (Smith).

–Les pronoms réflexifs ne sont pas lexicalisés, les deux lexèmes sont encore séparés : it self(e), my self(e), him self(e)), your selves “Take heed unto your selves”. Le dictionnaire permet d’enregistrer des mots composés. La correspondance entre les deux formes se fait donc simplement : it self,PRO+EN=itself. On peut le fléchir pour créer les deux formes équivalentes it self et it selfe. Ces entrées sont affectées du trait +UNAMB qui les désambiguïse.

–Les pronoms relatifs. Though peut s’orthographier tho’. Le pronom “which” peut être utilisé lorsque l’antécédent est humain, ce qui n’est plus le cas en AC: “The Georgians, which in some manner depend on the Greek Church, baptize not their children until they be eight years of age” (Rycaut).

–Morphologie verbale et paradigmes verbaux

L’adaptation des outils de flexion est requise, car les formes verbales correspondantes ne sont pas reconnues par les outils existants pour l’AC : quand il s’agit de flexion archaïque d’un lemme (verbe, pronom…) existant en AC, il faut construire un outil permettant d’ajouter la flexion manquante.

–Au présent, deux formes sont remarquables: la 2e personne du singulier, qui reçoit le morphème de flexion /est/ comme dans “thou satisfyest”, et la 3e personne du singulier qui prend le morphème flexionnel /(e)th/, comme dans “shee cometh”, “he hath”, “he doth”. Ces formes pouvant être rencontrées pour n’importe quel verbe, il faut les traiter dynamiquement par un graphe morphologique.

adviseth,advise,V+Tense=PR+Pers=3+Nb=s+EN=advise
desireth,desire,V+Tense=PR+Pers=3+Nb=s+EN=desire
saith,say,V+Tense=PR+Pers=3+Nb=s+EN=say
establisht,establish,V+Tense=PP+ EN=establish
establisht,establish,V+Tense=PT+Pers=1+Nb=p +EN=establish
fixt,fix,V+Tense=PP +EN=fix
fixt,fix,V+Tense=PT+Pers=1+Nb=p +EN=fix
linkt,link,V+Tense=PP +EN=link

Table 4 : Entrées ajoutées au dictionnaire pour le présent et le prétérit

–Au prétérit, certains verbes, qui sont irréguliers en AC, ont une forme fléchie en /ed/ (ex. : catched est le prétérit de catch, shined est le prétérit de shine), d’autres verbes dont l’infinitif se termine en p, k, x, ss, sh ont deux prétérits possibles, normal ou raccourci, ed devenant t ou ’t. D’autres verbes ont des formes différentes, par exemple to speak (I spake) or to begin (I begun, au lieu de began en AC). Ces formes seront insérées dans le dictionnaire.

L’apostrophe peut être utilisé pour le prétérit, le participe passé –voir la Figure 5–, ou l’adjectif déverbal qui prend alors la forme xyz’t ou xyz’d. Deux situations peuvent se présenter : xyz, forme tronquée, est néanmoins « reconnu » à tort (cas de establish’d), ou peut ne pas l’être (cas de judg’d). C’est un graphe syntaxique comportant les séquences <WF>’t ou <WF>’d qui complètera la reconnaissance. Notre corpus comporte 90 occurrences de ’d ou ’t.

Established se rencontre sous les trois formes : “and establish’d the same number”, “the establish’t doctrine”, “and for ever established the adoration”. Nous opérons la reconnaissance de ces formes au moyen d’un graphe morphologique, puis d’un graphe syntaxique, présentés

[46] *Nous offrons à toi de ce qui est à toi…*

Figure 5. Establish est faussement reconnu comme l'infinitif ou le présent. Le graphe syntaxique permet de construire established, qui est soit le participe passé, soit le prétérit, soit l'adjectif déverbal associés au verbe establish.

Étudions les cas de judg'd : "this being judg'd", la forme « judg » n'est pas identifiée par les dictionnaires. Nous appliquons les graphes morphologiques chargés de tester si moyennant certaines modifications, nous pouvons faire une hypothèse pour la reconnaissance de cette forme. La flexion ed, postposée à "judg" permet de construire judged le participe passé de judge. Le graphe syntaxique va confirmer cette interprétation et nous obtenons l'interprétation : judg,judge,PV+Tense=PP+ EN=judge.

Le mot professed est présent sous trois formes qui sont : professed, profes'd et profest : although "professed enemies to the Roman Church", "At the time of their being profest", "they believe Christianity can hardly be professed". Le cas de profest a été traité dans la Table 4. Il faut identifier « profes ». Le graphe ôtant un e muet émet la proposition profes,prof,N+Nb=p+Distribution=Hum+EN=prof, soit le pluriel de prof. Un autre graphe ajoute sed, et propose d'identifier professed, soit : profes,profess,PV+Tense=PP+EN=profess, (le graphe complet propose aussi de reconnaître le prétérit et l'adjectif déverbal). Par suite, le graphe syntaxique vérifie la présence de 'd et cela invalide la première possibilité, et valide la deuxième.

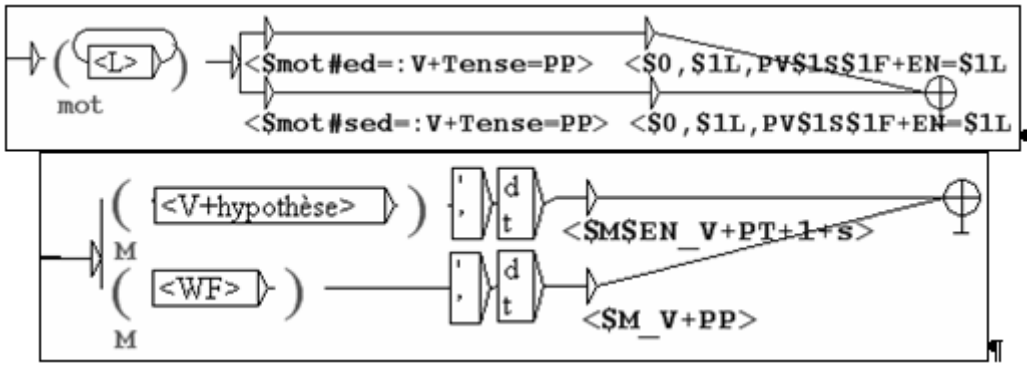


Figure 5 : Graphes permettant d'identifier le participe passé en 'd

Le cas de imbrac'd est différent car pour reconnaître embraced il faut effectuer les transformations : imbrac'd → embrac'd → embraced. Il faut donc combiner deux transformations.

–Prépositions et conjonctions. Though peut s'orthographier tho'. Through est parfois écrit thro' (e.g. "thrusting an iron stake thro' the body out under the neck"). Les prépositions sont souvent omises : "but our subject here is more tragical, the subversion of the sanctuaries of religion, the royal priesthood expelled their churches, et those converted into mosques" (ici from est omis).

–Déterminants : leur utilisation est très proche de celle de l'AC excepté pour l'article indéfini "an" devant des mots commençant par h comme dans "an holy amulet". On trouve aussi bien "an horrid impiety" que "a horrid sound". La transformation an → a est assez aisée car en AC on recense peu de mots commençant par un h non aspiré.

–"Wh-compounds" : les formes whosoever, wheresoever et whensoever sont couramment utilisées. Ces trois mots sont à ajouter au dictionnaire sous la forme: whosoever,PRO+EN=whoever.

4.2 Variations morphologiques et mots non reconnus

Il faut souligner l'aspect multilingue et religieux lié au contexte historique, géographique et culturel de ces textes, qui sont des récits de voyage, et à la propension des auteurs à recourir aux emprunts lexicaux que nous devons identifier et insérer avec les traits convenables à notre dictionnaire électronique. Notre corpus comporte des termes provenant de l'arabe, du grec, du latin, de l'italien et du turc.

4.2.1 ARCHAÏSMES, VARIATIONS SÉMANTIQUES ET ENTITÉS NOMMÉES

L'indexation d'un lexème du XVII^e^ siècle par les lemmes de l'AC peut conduire à différentes situations :

–Variation sémantique : les lexèmes sont reconnus, mais le sens a changé, ce sont les variations les moins flagrantes car le dictionnaire électronique pourra indexer faussement le mot qui ne sera pas repérable à ce niveau. Seul le lecteur averti peut repérer ce phénomène. Nous devons constituer un dictionnaire de tels termes et le rendre plus prioritaire que le dictionnaire anglais de NooJ. La reconnaissance prioritaire de penitentiary avec le sens de "spiritual father" va empêcher la comparaison de ce lexème aux dictionnaires moins prioritaires. Nous indiquons en Table 5, quelques entrées du dictionnaire, le code de flexion FLX=Nsp_y permettra de construire automatiquement le pluriel en remplaçant le y final par ies, soit penitentiaries.

penitentiary,N+FLX=Nsp_y+EN="spiritual father"+XVII
penitentiary,N+FLX=Nsp_y+EN="penitent"+XVII
ethnick,A+EN=pagan+ XVII

Table 5 : Exemples d'entrées de dictionnaire prioritaires

La priorité entre dictionnaires induit non seulement un ordre d'application, mais aussi des interdictions : les dictionnaires de niveau inférieur ne sont appliqués que si tous les dictionnaires de niveau de priorité supérieur n'ont pas produit de résultat.

–Le mot n'existe plus en anglais moderne, il est archaïque, ou bien le lexème étudié est la concaténation de plusieurs lexèmes en AC, il faut l'ajouter au dictionnaire.

–Le mot n'existe pas en anglais moderne parce qu'il s'agit d'une abréviation ou encore d'une Entité Nommée (EN) – par exemple d'un toponyme – or les ressources dictionnairiques ne comportent pas ou peu d'EN. Il faut les insérer dans le dictionnaire. Voir la Table 6.

Augustus,N+PR+m+s+Hum
Bajazid,N+PR+m+s+Hum+EN=Bajazet
Basileis Romaion,N+PR+p+Hum+EN="Emperor of the Greeks"
Nice,N+PR+s+Toponyme+EN=Nicæa
Romagnia,N+PR+s+Toponyme+EN="Greece and the Balkans"

Table 6 : Dictionaire de toponymes et de mots empruntés

4.2.2 VARIATIONS MORPHOLOGIQUES

Le mot peut avoir subi une transformation morphologique comme le redoublement de lettres, l'insertion/suppression du e muet, ou une modification comme le remplacement de /edge/ par /ege/. Ce sont des transformations morphologiques dont nous pouvons décrire le paradigme, et traiter ces cas par un transducteur qui identifie, teste et traite le lexème, construit le lemme moderne correspondant et propose une entrée pour le dictionnaire.

Certaines des modifications peuvent se retrouver aussi bien à l'intérieur du mot qu'au début ou à la fin. Dans ce cas, dans le paradigme, soit X, soit Y est vide. Nous présentons quelques exemples en Table 7, et la Figure 6 présente des graphes de transformation..

Modification	**Exemples**	**Paradigme**
im pour em	imbrace pour embrace	XimY → XemY
en pour in	encreasing pour increasing	XenY→ XinY
in pour en	Intangling, intrench pour entangling, entrench	XinY → XenY
ncy pour nce	occurrency pour occurrence	XncyY → XnceY
ous pour ate	degenerous pour degenerate	XousY → XateY
ick pour ic	Arabick, garlick pour Arabic, garlic	XickY → XicY
ik pour ic	traffik pour traffic	XikY → XicY
ie pour y	christianitie, pour christianity	XieY → XyY
ie, ey pour y	countrey pour country	XeyY → XyY
our pour or	emperour, terrour pour emperor, terror	XourY → XorY
oa pour o	shoar, cloaths, choake	XoaY → XoY
edge pour ege	alledge, colledge pour allege, college	XedgeY →XegeY
ai pour ei	soveraigne	XaiY → XeiY
ea pour e	compleate, seaven	XeaY → XeY
ph pour f	phantastique	XphY → XfY
w pour u	perswasion pour persuasion	XwY → XuY
y pour i	oyl, coyn pour oil, coin	XyY → XiY
i pour j	iourney pour journey, Iew pour Jew	XiY → XjY
eer pour ear	yeer, neer pour year, near	XeerY → XearY
ence pour ense	expence	XenceY → XenseY

Table 7 : Exemples de modifications morphologiques

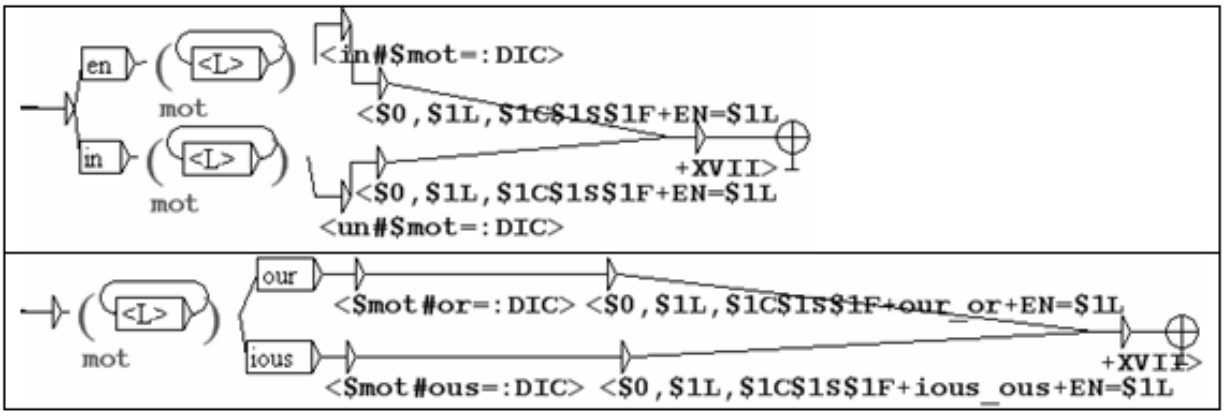


Figure 6 : Graphes de transformations morphologiques

L'application de ces graphes permet d'identifier encreasing comme l'adjectif et le nom increasing, ou le verbe increase, et governours comme le nom pluriel governor. La Table 8 présente des entrées du dictionnaire. Le trait +XVII marque la diachronie. Le mot identifié comme étant en anglais du XVII^{e} est suivi de sa forme moderne, de sa catégorie codée et de traits syntaxiques, le trait EN=XXX indique que l'équivalent en AC est XXX.

encreasing,increasing,A+EN=increasing+XVII
encreasing,increase,N+Nb=s+EN=increase+XVII
encreasing,increase,V+Tense=G+EN=increase+XVII
governours,governor,N+Nb=p+Distribution=Hum+EN=governor+XVII
inferiour,inferior,N+Nb=s+Distribution=Hum+EN=inferior+XVII
inferiour,inferior,A+EN=inferior+XVII

Table 8 : Entrées proposées par le graphe

4.3 Traitement de formes discontinues : reconnaissance et règles de réécriture

–Le subjonctif (qui est formé grâce au verbe à l'infinitif sans to à toutes les personnes) est très commun. Il apparaît dans les clauses conditionnelles et concessives (contrairement à l'AC où il est utilisé principalement dans des contextes formels). Il est précédé de unless | lest | if | provided | though | whether | whatever etc., comme dans "provided it be done after a due manner", "if they be kaloirs" (Smith), "whatever it be" (Smith), mais aussi dans des clauses temporelles, "until they be eight years of age" (Rycaut), et dans les clauses comparatives, "they had rather remaine as they be" (Sandys). Le graphe présenté en Figure 7 réécrit "provided it be" en "provided it is" et "if the criminal be" en "if the criminal is".

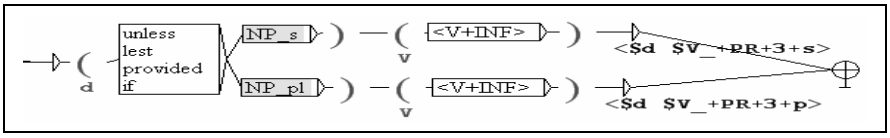


Figure 7 : Graphe de reconnaissance du subjonctif

–Formes avec soever

Les structures nominales suivantes sont remarquables: no <N0> of what <N1> + soever peut se traduire whatever the N1 of a N0. On trouve la structure « *at what great distance soever* » (Smith, 186), qui donnerait *« however great the distance is »* en anglais moderne, ou « *of what communion soever* », « of *whatever communion »* en anglais moderne, de même que la construction adjectivale similaire : *how <A> soever* se transpose en " *however adjectif* " *:* "*How strict soever this Church is esteemed in admitting many degrees of marriage*" (Rycaut). La Figure 8 présente ces graphes de reconnaissance et de réécriture de telles séquences en AC.

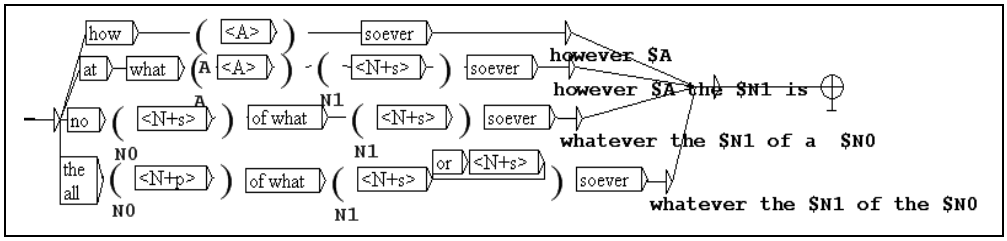


Figure 8 : Graphe de reconnaissance des formes "soever"

–Formes déclarative, interrogative et négative.

À la forme déclarative du présent et du prétérit, à côté de la forme usuelle du verbe (avec la flexion "s" à la troisième personne du singulier), l'auxiliaire do(e) peut être utilisé avec l'infinitif: "They [the Greeks] are of diverse trades in cities, and in the country do till

(=cultiver) the earth" (Sandys). "Mens minds did labour with fearefull expectations" (Sandys).

Nous pouvons traiter ces formes par le graphe présenté en Figure 9. Il reconnaît les séquences (do | doe | did | does | doth) [ADV] V. L'adverbe est optionnel. L'auxiliaire do est supprimé et le verbe est mis au temps voulu : lemme, prétérit, ou troisième personne du singulier du présent. Nous indiquons quelques exemples de résultats : doe still remain/<still remain>, doth believe them/<believes them> et did sit/<sat>.

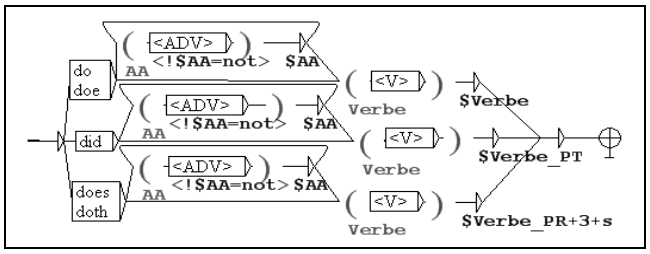


Figure 9 : Graphe de transcription de la forme déclarative

–Do est aussi utilisé à l'impératif: "do thou bless this oblation", "do thou pardon, as being our good and gracious God" (Smith). Un graphe permet de transcrire cette forme en AC en remplaçant do thou <V> par <V>.

–À la forme négative, à côté de l'auxiliaire "do" + not + verbe, on trouve aussi la structure verbe + not, comme dans "the women marry not till the age of 24" (Sandys), ou dans la phrase "their difference theologicall I enquired not" (Blount).

–À la forme interrogative, le verbe avec un sujet inversé est parfois utilisé. Dans la langue du XVII^e siècle, les questions ne nécessitent pas toujours l'auxiliaire "do". Donnons quelques exemples tirés de deux pièces d'Aphra Behn : "What mean you by this language?" (*The Dutch Lover*, 1673). "How came she in?" (*Abdelazer or the Moor's Revenge*, 1676). Nous n'avons pas traité ces formes absentes de notre corpus.

4.4 Évaluation

Notre travail a été construit à partir d'un corpus de textes représentant 44300 mots environ. Nous présentons l'évaluation de notre traitement sur les textes des six auteurs sélectionnés. La Table 9 fournit un nombre de mots inconnus, ainsi que de leurs occurrences, pour chacun des six auteurs.

Auteur	Année	Taille du corpus	Mots différents	Mots inconnus	occurrences des mots inconnus
Blount	1636	1408	621	132	190
Sandys	1652	3910	1419	310	427
Rycaut	1679	12920	2270	345	656
Smith	1682	22640	4285	710	1266
Wheler	1682	2365	747	80	127
Ray	1693	1140	533	60	79

Table 9 : Évaluation des mots inconnus dans le corpus traité

Nous avons évalué les pourcentages des mots inconnus par rapport au nombre de mots différents et par rapport au corpus. Ces pourcentages sont en moyenne respectivement de 16,58 % et de 6,18 %. Nous avons appliqué nos dictionnaires et nos graphes à ces mots

inconnus et nous évaluons les résultats dans la Table 10, en distinguant les mots étrangers, les noms propres, et les abréviations des mots du dix-septième siècle. Nous constatons que ce dernier pourcentage est plus grand pour les textes les plus anciens. Ces mots sont repérés soit par un dictionnaire *ad hoc*, soit par un graphe, les résultats que nous donnons ont été validés. Les formes ainsi reconnues par un graphe sont insérées dans les dictionnaires.

Auteur	Mots inconnus	Langue étrangère	Noms propres	Abréviations	XVII^{e}
Blount	132	0,70%	7,48%	0,47%	91,36%
Sandys	310	3,43%	12,15%	0,66%	82,56%
Rycaut	345	21,25%	14,81%	0,87%	60,98%
Smith	710	13,69%	18,75%	0,60%	63,91%
Wheler	80	12,50%	35,58%	2,88%	50,96%
Ray	60	13,40%	26,80%	2,06%	58,76%

Table 10 : Catégorie des mots inconnus

Parmi les mots non reconnus nous devons faire un traitement particulier pour les formes autrefois écrites avec un tiret. En effet sur 128 formes x-y telles que 'mid-land' présentes dans les textes, nous avons identifié 43 formes aujourd'hui écrites xy comme 'midland'. Celles-ci sont également insérées dans le dictionnaire.

En ce qui concerne les erreurs, leur nombre est faible, ce qui s'explique par le fait que nos graphes et dictionnaires ont été construits pour traiter ce corpus. C'est sur le traitement d'autres corpus que pourrait se faire l'évaluation de notre travail.

Enfin certains problèmes sont liées au plurilinguisme du texte : in peut être du latin in nomine Patris, ou de l'italien Che fida in Grego, to peut être du grec : To katasphragisai to paidion. Les mots empruntés, généralement non reconnus, vont être traités, à tort, par les graphes, et vont augmenter le nombre d'erreurs. Il conviendrait d'encadrer les citations de marques permettant d'appliquer les dictionnaires adéquats.

Nous donnons ici un exemple extrait des textes traités, et indiquons en caractères gras les réécritures effectuées : … when Antipater, Perdiccas, Seleucus, Lysimachus, Antigonus, Ptolemey, and the rest of the successors of Alexander had shared his empire among them, they endeavoured as much as they could to plant his new-got **kingdomes/ kingdoms** with their **countrey-men/ countrymen**: whose posterity in part **remaineth/ remains** to this day, (though **vassaled/submitted** to the often changes of **forain/ foreign governours/governors: supplyed/ supplied** by the **extention/extension** of the latter Greek Empire; who yet retain wheresoever/**wherever** they live, their name, their religion, and particular language. A nation once so excellent, that their precepts and examples **doe/do** still remain as approved canons to direct the mind that **endevoureth/endeavours vertue/ virtue**. Admirable in arts and glorious in arms; famous for government, **affectors/lovers** of **freedome/ freedom**, every way noble: and to whom the rest of the world were reputed barbarians.

5 CONCLUSION

La partie automatisable de notre travail, même si elle apporte une aide précieuse, ne peut se concevoir que comme une étape suivie d'une validation. Les aspects grammaticaux ne sont actuellement pas tous traités. Si nous sommes en mesure de résoudre dans des proportions significatives l'insertion des apostrophes, et la réécriture de certaines formes archaïques, nous n'avons pas encore de moyen fiable de reconnaissance automatique des adjectifs postposés, l'utilisation d'adjectifs comme adverbes, et l'omission de prépositions. Le repérage des

prépositions omises suppose un lexique-grammaire des verbes, le repérage de l'antécédent du relatif which pose problème, car il nécessite l'identification de traits humains et du genre dans les dictionnaires ce qui est un très gros travail. Enfin nous devons souligner que la combinaison des modifications que nous décrivons pose des problèmes autrement plus ardus que leur traitement séparé.

Dans ce travail la partie lexicographique est la plus aisée à mener à son terme. En élargissant le corpus et en créant une base de données consacrée aux récits de voyage anglais au XVII^e^ siècle, nous voulons créer un dictionnaire qui inclut les archaïsmes et les mots étrangers et peut donner lieu à des éditions spécifiques. Le choix de traits sémantiques facilite l'établissement de glossaires spécialisés : glossaire religieux, ou de nature historique ou géographique. Lors de la recherche en ligne sur catalogue informatisé, le recensement des variantes orthographiques d'un même mot permettrait de retrouver des ouvrages en tapant un titre avec une orthographe moderne.

L'existence d'un dictionnaire de la langue anglaise du XVII^e^ siècle serait d'un grand intérêt pédagogique. Il permettrait de recenser les principales particularités sémantiques et morphosyntaxiques de l'anglais du XVII^e^ siècle et de proposer des équivalents en anglais moderne. Deviennent aisément lisibles et abordables par nos étudiants et par le grand public (anglophone ou non) des textes jusque-là oubliés ou ensevelis dans les Rare Books Rooms des bibliothèques anglaises ou accessibles seulement aux chercheurs dûment accrédités. Ajoutons que le corpus va s'élargir au fil du temps puisque nous travaillons désormais sur d'autres aires géographiques.

Nous espérons ainsi rendre les textes du XVII^e^ siècle, et en particulier les récits de voyage, qui sont une source historique précieuse, plus accessibles aux étudiants non anglophones et non-spécialistes et leur éviter de recourir aux traductions de l'époque qui non seulement sont de belles infidèles mais omettent parfois de traduire des phrases ou des passages entiers du texte original.

6 RÉFÉRENCES

Blount H. (1636). *A Voyage into the Levant*. London.

EEBO (Early English Books Online) : http://eebo.chadwyck.com/home

Freeborn D. (1998). *From Old English to Standard English. 2nd ed. Basingstoke*. Palgrave.

Hogg R. M., Blake N. F., Lass R., Romaine S., Burchfield R. W. et Algeo J. (1992). *The Cambridge History of the English Language. Vol. III*. Cambridge : CUP.

Michael I. (1987). *The Teaching of English: From the Sixteenth Century to 1870.* Cambridge : CUP.

Pignot H. (2006). « Comparaison entre l'anglais du XVII^e^ siècle et l'anglais britannique moderne à travers les récits de trois voyageurs britanniques ». *Colloque NooJ-Trad*, Paris, 26 septembre 2006.

Pignot H. (2007). *La Turquie chrétienne: récits des voyageurs français et anglais dans l'Empire ottoman au XVII^e^ siècle.* Vevey, Xénia.

Pignot H. (2009). *Christians under the Ottoman Turks: French and English Travellers in Greece and Anatolia (1615-1695).* Piscataway: Gorgias Press.

Pignot H. et Piton O. (2008). « Language processing of 17th Century British English with NooJ », *Workshop NooJ 2008*, Budapest, juin 2008, à paraître.

Piton O. et Lagji Kl. (2008). « Morphological study of Albanian words, and processing with NooJ ». Dans X. Blanco et M. Silberztein (éds.), *Proceedings of the 2007 International NooJ Conference*, Cambridge: CSP, 2008.

Ray J. (1693). *A Collection of Curious Travels and Voyages through the Levant, tome II.* Leiden.

Rycaut P. (1679) *The Present State of the Greek and Armenian Churches.* London.

Sandys G.(1652). *A Relation of a Journey begun An. Dom. 1610 (-1615).* London.

Silberztein M. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX.* Paris : Masson.
Silberztein M. (2005). « NooJ's dictionaries ». Dans Zygmunt Vetulani (éd.), *Proceedings of the 2nd Language & Technology Conference*, April 21-23, 2005, Poznán, Poland.
Smith Th. (1680). *An Account of the Greek Church.* London.
Wheler G. (1682). *A Journey into Greece.* London.

VERBES INTENSIFIEURS : UNE RECHERCHE QUALITATIVE ET QUANTITATIVE À PARTIR DU WEB FRANCOPHONE

Ewa Pilecka
Institut d'Etudes Romanes - Université de Varsovie
pilecka@hotmail.com

RÉSUMÉ

L'article ci-dessous est consacré à la présentation d'une recherche menée à partir du corpus des textes du Web francophone et ayant pour but de répertorier et décrire les paradigmes de verbes et de noms qui peuvent former la collocation intensive *V de N* de type *rougir de colère*, *mourir d'ennui*, *rayonner de bonheur*. Une recherche sur corpus permet de décrire certaines caractéristiques de cette construction que d'autres méthodologies laissent de côté. Nous montrerons également comment contourner les difficultés posées par la recherche à objectif linguistique menée à l'aide d'un moteur de recherche d'accès public.

1 INTRODUCTION

Les procédures de recherche en linguistique s'articulent entre deux pôles constitués, d'une part, par l'introspection, et de l'autre, par la recherche sur corpus ; entre les deux, on a le vaste champ de l'exemple recueilli soit dans les textes, soit sur le terrain. Les préférences pour l'un ou l'autre sont bien sûr chose individuelle. Pour un linguiste n'étant pas locuteur natif, le recours au « plus vaste corpus jamais constitué » (à savoir, celui des textes du Web) peut être une tentative irrésistible ; il faut cependant se rendre compte du fait qu'à côté de ses avantages indéniables[47] qui sont l'accessibilité, la gratuité, la rapidité d'interrogation, la diversité des genres, le caractère parfaitement synchronique (vu ces deux dernier points, on peut même avancer la thèse que « le Web est la langue »), il y a un certain nombre de difficultés à contourner, tant d'ordre méthodologique que purement technique. Nous en parlerons aussi à l'exemple de notre recherche concernant les verbes intensifieurs du français, menée à partir du corpus des textes du Web francophone.

2 LA NOTION DE VERBE INTENSIFIEUR

Les verbes que nous appelons « intensifieurs » apparaissent dans la construction *V de N* décrite par Leeman 1991 (p.ex. *rougir de colère*, *trembler de peur*, *pleurer de bonheur, mourir d'ennui*, *rayonner de joie*...) et correspondent à deux parmi les trois sous-classes y examinées, à savoir celles ou *de N* est:

1. supprimable mais difficilement déplaçable (rougir de confusion, crier d'excitation, tituber de fatigue);
2. ni supprimable, ni déplaçable (crouler de rire, bouillonner de rage, brûler d'amour).

[47] A ce sujet, cf. Wooldridge (2005).

Dans ce type de collocations c'est le substantif qui assume le rôle du prédicat central, et le verbe en indique le degré d'intensité (en l'occurrence, il s'agira toujours d'un degré d'intensité élevé).[48]

La sous-classe où le verbe est supprimable et déplaçable, voire meilleur en tête de la phrase se situe en dehors de notre champ d'investigation, car le verbe a alors a la fonction du prédicat central indiquant la conséquence d'un affect ou d'un état psychologique ou physiologique, exprimé par le nom qui a la fonction du complément circonstanciel de cause. C'est de ce schème syntaxico-sémantique primaire, dont est dérivée - par voie de schématisation et grammaticalisation successive – la classe des collocations à valeur intensive.

Le cadre syntaxique *V de N*, dans lequel apparaissent les verbes intensifieurs, est peu spécifique[49]. Leur caractère fonctionnel est dû aux les restrictions de sélection associés à cette structure et qui concernent aussi bien les substantifs intensifiés que les verbes intensifieurs.

Comme nous l'avons pu constater, les grammaires et les dictionnaires traditionnels (p.ex. Nouveau Petit Robert, Trésor de la Langue Française) passent le plus souvent sous silence la fonction intensificative de ces verbes; sauf quelques cas manifestes (*mourir*, *rougir*), la définition lexicographique n'indique pas la possibilité de marquer l'intensité élevée d'un sentiment, d'une sensation etc. à travers la construction *V de N*. De même, les entrées correspondant aux N intensifiables (qui correspondent à plusieurs sous-classes[50] des « noms intensifs » de Flaux, Van de Velde 2000) ne le signalent pas de manière systématique, et lorsqu'ils le font, la valeur intensive des collocations n'est qu'exceptionnellement explicitée (qui est cependant confirmée par les paraphrases proposées dans Zingle 2003, p.ex.:

- *mourir de rire* = 'rire beaucoup'
- *trembler de colère* = 'être très en colère'
- *mourir de nostalgie* = 'éprouver une profonde nostalgie'
- *trembler de fièvre* = 'avoir une forte fièvre'
- *trembler de joie* = 'éprouver une grande joie'
- *trembler d'émotion* = 'éprouver une vive émotion'

Cependant, même les dictionnaires de collocations (Beauchesne 2002, Zingle 2003) sont loin d'être exhaustifs, et nous avons pu, *a posteriori*, y constater l'absence de bon nombre de collocations que nous avons recensées à partir du Web, avec, parfois, un nombre d'occurrences significatif (comme p.ex. *pleurer de soulagement* - 45 occurrences, *crier d'extase* - 40 occurrences, *se pâmer d'extase* - 33 occurrences, *soupirer de lassitude* - 30 occurrences, *baver de convoitise* - 25 occurrences, *fondre d'attendrissement* - 19 occurrences, *rosir de confusion* – 14 occurrences, etc).

---

[48] Leeman 1991 fournit une centaine d'exemples de collocations *V de N* qui correspondent à cette caractéristique, et qui, soumis à l'évaluation d'un échantillon de locuteurs natifs, ont été acceptées par au moins 150 personnes.

[49] Cf. *parler de sa vie, s'approcher de qq, se servir d'un marteau, vivre d'amour et d'eau fraîche...* , où le verbe n'est pas un intensifieur ; le syntagme prépositionnel en *de* qui l'accompagne peut avoir diverses fonctions (complément d'objet indirect, complément locatif, divers circonstanciels...).

[50] La classe des noms abstraits intensifs comporte les noms de qualités, les noms d'affects (sentiments - prédicats à deux arguments ou émotions - prédicats à un argument) et les noms d'états (psychologiques ou physiologiques). Le cadre syntaxique *V de N* se prête principalement à l'intensification des noms d'affects et des noms d'états . Il s'agit des noms prédicatifs à argument humain, qui dénotent une expérience intérieure, subjective, psychique ou physique ; nous les appelons « noms d'expérience intérieure ». Exceptionnellement, on y trouve également des exemples de noms de qualités, p.ex. :

Je passe sur le contenu des articles qui sont à vomir de médisance!
Ta lettre me fait sursauter de beauté, chaque fois que je la lis.

3 RECHERCHE QUALITATIVE

L'objectif premier de notre étude était d'ordre qualitatif: il s'agissait de dresser une liste (exhaustive, dans la mesure du possible) des verbes pouvant jouer le rôle d'intensifieur. Ensuite, à partir des collocations intensives *V de N* fonctionnant en français, nous avons prévu d'examiner les facteurs sémantico-cognitifs qui fondent l'association du verbe intensifieur avec le substantif intensifié.

Le corpus dépouillé était celui des textes du Web francophone, consulté du 10.06.08 au 02.08.08, à l'aide du moteur de recherche Google.be. La recherche a été limitée à la construction intransitive et pronominale; nous en avons donc exclu:

15. les constructions transitives (aussi bien celles à verbe support que celles avec un verbe à sens plein), p.ex.: pousser un cri, une exclamation, un gémissement, un ouf, un soupir; verser une larme...; hocher la tête, lever les sourcils, grincer les dents, se frotter les mains, serrer les poings ....;
16. les constructions à verbe copule accompagné d'un adverbe, d'un adjectif ou d'un participe passé en emploi adjectival: se trouver mal, rester muet, devenir rouge, être fou....;
17. les constructions qui n'apparaissent que dans la forme négative: *ne plus se sentir de, ne pas se posséder de* etc.

La recherche des collocations s'est faite « en spirale ». Lors d'une première étape, nous avons vérifié la présence et le nombre d'occurrences des collocations figurant sur la liste de Leeman 1991 (à notre grande surprise , certaines de ces collocations, n'ont pas trouvé d'attestation sur les pages francophones du Web). La plupart des noms intensifiés appartenant à la classe des « noms de sentiment », nous avons ensuite élargi notre recherche aux substantifs de deux autres listes (Mathieu 2001, Badynska-Lipowczan 1999), au total plus de 400 items différents. Afin de détecter leurs collocatifs verbaux, nous avons testé plusieurs contextes « à joker » et avons finalement choisi comme la plus opératoire la suite « a * de N » / « est * de N » (N étant la variable donnée, et le joker * - la variable recherchée, à savoir le verbe sous la forme de participe passé). Nous avons pu dresser ainsi une liste de plus de 200 verbes intensifieurs que nous avons groupés en plusieurs sous-classes sémantiques:

- mouvements : trembler, palpiter, se contorsionner, tomber, sursauter...
- immobilité : se figer, se pétrifier, s'engourdir....
- activités : applaudir, danser, se cacher, se défénestrer...
- parole/paraverbal : rugir, hurler, glousser, bredouiller, s'exclamer, siffler, chantonner....
- maladie/mort : mourir, crever, décéder, défaillir, s'évanouir...
- changement de couleur : blanchir, rougir, verdir, blêmir...
- respiration : s'étouffer, soupirer...
- secrétions : suer, baver, pisser...
- autres physiologiques : renifler, peler, bander...
- température : brûler, bouillir, se glacer...
- métaphore du contenant : déborder, se gonfler, éclater...
- métaphore de la lumière : rayonner, briller, irradier...

4 RECHERCHE QUANTITATIVE

La seconde étape avait pour but de chercher les associations préférentielles entre les V et les N. Nous avons recouru au contexte fixe « V de * » / « V d * », afin de détecter les substantifs apparaissant avec tel ou tel verbe.

Confrontée au choix entre la recherche d'une suite de mots exacte et celle de formes lemmatisées, nous avons opté pour la première solution; nous avons décidé de limiter la recherche à la forme canonique *V inf de N* , car l'infinitif:

22. est très rarement homographe avec d'autres formes (la seule exception, en ce qui concerne les verbes susceptibles d'entrer dans notre corpus, était l'infinitif des verbes *rire* et *sourire*) ce qui n'est pas le cas p.ex. du participe passé[51] ;
23. n'est pas susceptible de varier en nombre et genre (sauf pour les verbes pronominaux, dont le pronom peut avoir différentes formes) ni avoir des variantes morphologiques du radical en fonction du temps et du mode (ce qui est le cas des formes finies);
24. n'exclut pas l'apparition, dans la phrase, de tout un éventail de valeurs temporelles[52] et modales, car ce cadre syntaxique n'en interdit, *a priori*, aucune.

Deuxième constatation surprenante: même s'il existe, pour la plupart des substantifs, des associations préférentielles, le choix du verbe intensifieur est loin d'être restreint. Nous avons également relevé des collocations explicitement exclues par Leeman (p.ex. *rayonner de honte*), ainsi que des collocations que nous-même considérions comme peu probables.

Les premiers résultats de la recherche se sont avérés prometteurs. Cependant, une recherche linguistique menée à l'aide d'un outil informatisé dont ce n'est pas l'objectif principal rencontre inévitablement un certain nombre de problèmes méthodologiques, notamment lorsqu'il était question d'évaluer le nombre d'occurrences d'une collocation donnée. A cette étape, nous avons eu recours à une vérification manuelle et semi-manuelle[53] systématique. Elle avait pour but d'éliminer:

52. les bruits provenant de la traduction automatique;
53. les entrées des dictionnaires en ligne;
54. les pages réitérés;
55. les exemples manifestement incorrects, produits par des locuteurs non-francophones[54];
56. *les collocations associés de commentaires métalinguistiques,* en particulier celles signalées comme incorrectes.

Le temps de recherche prévu initialement devait être relativement court afin d'éviter les difficultés posées par l'accroissement constant du corpus de textes. Cependant, l'ampleur inattendue des résultats nous a amené à consulter le corpus Web à plusieurs reprises, aux dates différentes, en l'espace de 10 mois, ce qui nous a forcé de chercher les moyens de standardisation des résultats obtenus aux étapes successives de la recherche (grâce aux fonctions de recherche avancée proposée par Google ainsi qu'aux opérations statistiques, p.ex. extrapolation des données à partir d'un échantillon, vérification des fluctuations observées dans le corpus, calcul de la marge d'erreur, passage des chiffres absolus aux tranches de centiles).

Le nombre et la diversité des associations a dépassé nos attentes: nous avons recensé, au total, 168 substantifs intensifiables et 237 verbes intensifieurs susceptibles d'entrer en

[51] Le moteur de recherche ne faisant pas la distinction entre les caractères avec et sans signes diacritiques, le participe passé des verbes du 1er groupe de conjugaison se confond facilement avec le présent (*chante* / *chanté*)

[52] Notamment, lorsqu'on a affaire à la construction factitive *faire + Vinf.*

[53] Notamment pour l'élimination des pages réitérées.

[54] Relativement peu nombreux, présents essentiellement dans les discussions/commentaires des internautes.

collocation avec eux, ainsi que 2651 collocations effectivement employées[55] représentées par un nombre d'occurrences bien diversifié (de 1 à >500 occurrences[56]).

Le nombre des noms intensifiés par un verbe donné varie en fonction de ce dernier. En tête de la liste figure le verbe *(se) mourir*, qui entre en collocation avec 86 noms (les chiffres à côté de chaque nom indiquent le nombre d'occurrences relevées dans le Web):

(se) mourir + 86 noms:

amour >500, *chagrin* >500, *ennui* >500, *envie* >500, *épuisement* >500, *faim* >500, *froid* >500, *honte* >500, *peur* >500, *plaisir* >500, *rire* >500, *soif* >500, *inanition* 451, *douleur* 277, *jalousie* 246, *joie* 232, *fatigue* 207, *désespoir* 190, *bonheur* 189, *tristesse* 189, *chaleur* 171, *impatience* 170, *solitude* 155, *dépit* 91, *effroi* 90, *trouille* 87, *désir* 86, *rage* 77, *stress* 65, *inquiétude* 64, *colère* 61, *extase* 55, *frayeur* 49, *terreur* 42, *regret* 36, *sommeil* 34, *lassitude* 32, *dégoût* 26, *fièvre* 26, *souffrance* 24, *confusion* 18, *surprise* 18, *faiblesse* 17, *indignation* 17, *remords* 17, *émotion* 16, *mélancolie* 16, *volupté* 16, *curiosité* 15, *épouvante* 14, *ivresse* 13, *culpabilité* 11, *orgueil* 10, *pitié* 10, *humiliation* 9, *nostalgie* 9, *tendresse* 9, *anxiété* 7, *gêne* 7, *malheur* 7, *admiration* 6, *béatitude* 6, *amertume* 5, *énervement* 5, *passion* 5, *compassion* 4, *écoeurement* 4, *exaspération* 4, *incertitude* 4, *reconnaissance* 4, *aise* 3, *contentement* 3, *douceur* 3, *émoi* 3, *satisfaction 3, attendrissement* 2, *découragement* 2, *éblouissement* 2, *fierté* 2, *stupeur* 2, *timidité* 2, *bien-être* 1, *gratitude* 1, *hébétude* 1, *mépris* 1, *rancune* 1;

viennent ensuite:

pleurer + 71 noms:

bonheur >500, *émotion* >500, *joie* >500, *rage* >500, *rire* >500, *douleur* 282, *honte* 268, *désespoir* 237, *tristesse* 182, *amour* 160, *peur* 130, *jalousie* 103, *plaisir* 95, *dépit* 86, *chagrin* 74, *colère* 65, *nostalgie* 59, *admiration* 58, *attendrissement* 56, *faim* 56, *fatigue* 55, *déception* 53, *frustration* 52, *pitié* 48, *énervement* 46, *soulagement* 45, *compassion* 40, *angoisse* 36, *dégoût* 36, *épuisement* 25, *peine* 24, *regret* 23, *haine* 22, *reconnaissance* 22, *gratitude* 21, *découragement* 20, *extase* 18, *indignation* 18, *solitude* 17, *amertume* 16, *humiliation* 16, *terreur* 16, *émerveillement* 13, *fierté* 13, *remords* 13, *trouille* 13, *culpabilité* 10, *aise* 9, *contentement* 8, *douceur* 8, *confusion* 6, *écoeurement* 6, *soif* 6, *désir* 5, *épouvante* 5, *lassitude* 5, *malheur* 5, *mélancolie* 5, *bien-être* 4, *exaspération* 4, *fureur* 4, *tendresse* 4, *allégresse* 3, *éblouissement* 3, *volupté* 3, *incertitude* 2, *appréhension* 1, *béatitude* 1, *convoitise* 1, *désapprobation* 1, *irritation* 1;

frémir + 61 noms:

horreur >500, *plaisir* >500, *peur* 205, *impatience* 154, *bonheur* 145, *joie* 109, *effroi* 102, *indignation* 93, *angoisse* 92, *aise* 90, *terreur* 79, *émotion* 59, *colère* 54, *excitation* 54, *dégoût* 49, *rage* 49, *inquiétude* 38, *crainte* 36, *épouvante* 36, *amour* 30, *honte* 30, *enthousiasme* 29, *extase* 22, *trouille* 17, *frayeur* 13, *froid* 13, *allégresse* 12, *appréhension* 12, *volupté* 12, *pitié* 11, *espoir* 10, *nostalgie* 10, *orgueil* 10, *passion* 10, *bien-être* 8, *espérance* 8, *émoi* 7, *ennui* 7, *haine* 7, *fureur* 6, *anxiété* 5, *curiosité* 5, *surprise* 4, *incertitude* 3, *mépris* 3, *tendresse* 3, *convoitise* 2, *émerveillement* 2, *faiblesse* 2, *gêne* 2, *malheur* 2, *mélancolie* 2, *pudeur* 2, *écœurement* 1, *exaspération* 1, *gratitude* 1, *hésitation* 1, *humiliation* 1, *malaise* 1, *regret* 1, *remords* 1;

trembler + 61 noms:

peur >500, froid 280, *effroi* 203, *émotion* 157, plaisir 142, joie 81, terreur 77, *colère* 70, *frayeur* 66, *bonheur* 52, *horreur* 49, excitation 45, désir 43, *impatience* 43, trouille 42, crainte 41, *épouvante* 39, *inquiétude* 37, angoisse 34, amour 32, émoi 31, *indignation* 30, *fièvre* 27, *appréhension* 19, *honte* 17, admiration 13, *énervement* 13, *fureur* 13, *fatigue* 11, haine 11,

[55] Soit environ 6,5% du nombre de collocations théoriquement possibles (237x168=39816).

[56] Les collocations les plus fréquentes ont sans aucun doute des milliers d'occurrences; cependant, comme le moteur de recherche indique le nombre exact de pages contenant la suite de mots recherchée lorsque celui-ci est inférieur à 1000, et vu la présence de pages à contenu quasi-identique, nous avons trouvé utile de fixer la « limite de consultabilité » à 500 occurrences.

anxiété 9, espoir 9, *extase* 9, *aise* 8, faim 6, *dégoût* 4, enthousiasme 4, *incertitude* 4, stupeur 4, surprise 4, *convoitise* 3, frustration 3, *pitié* 3, timidité 3, curiosité 2, *ennui* 2, *épuisement* 2, *fierté* 2, *gratitude* 2, hésitation 2, *passion* 2, *trac* 2, *concupiscence* 1, *confusion* 1, émerveillement 1, *exaspération* 1, humiliation 1, *irritation* 1, orgueil 1, remords 1, solitude 1.

Dans le cas de la plupart des verbes, on a affaire à des associations préférentielles, mais non figées. Une minorité de substantifs (33 items) ne s'accompagnent que d'un seul verbe intensifieur. Leur nombre d'occurrences est également variable, p.ex.:

danser de joie - 325 occurrences

geler de froid - 58 occurrences

peler de froid - 46 occurrences

...

roter de satisfaction - 7 occurrences

renifler de mépris – 6 occurrences

...

puer d'orgueil – 1 occurrence

caracoler d'aise - 1 occurrence

Il est intéressant de remarquer que *puer d'orgueil* (une seule occurrence dans notre corpus), et *caracoler de joie* (dont nous n'avons recensé aucune occurrence) figurent parmi les collocations intensives acceptées « sans hésitation » par le groupe de locuteurs natifs consulté par Leeman (tandis que *caracoler d'aise* n'y est pas mentionné). Cela prouve que l'acceptabilité d'une collocation (donc, son caractère facilement interprétable) et son utilisation effective par les locuteurs sont deux choses différentes.

Nous avons dressé les listes de verbes intensifieurs pour chaque substantif intensifiable; là aussi, les paradigmes diffèrent entre eux quant au nombre d'items qu'ils contiennent[57], certains noms se laissent intensifier par un grand nombre de verbes, p.ex.

***joie* + 88 verbes:**

applaudir, bafouiller, bander, *barrir*, baver, *bêler*, *beugler*, bondir, bouillir, *bouillonner, bourdonner*, brailler, chanceler, chanter, chavirer, chialer, convulser, *crépiter*, crever, crier, danser, déborder, se défenestrer, délirer, éclater, s'écrouler, s'effondrer, s'envoler, *éructer*, s'étrangler, s'évanouir, s'exclamer, exploser, *flamboyer*, fondre, fredonner, frémir, frétiller, frissonner, gambader, glousser, *grimacer*, grogner, *hennir*, hurler, s'illuminer, imploser, irradier, *miauler*, mourir, *mugir*, palpiter, (se) pâmer, se pendre, pépier, périr, pétiller, piaffer, piétiner, pleurer, postillonner, *rayonner*, respirer, *resplendir*, rosir, rougir, rugir, *rutiler,* sangloter, sauter, sautiller, siffloter, souffler, soupirer, sursauter, tituber, *tomber,* toussoter, transpirer, trembler, se trémousser, trépigner, tressaillir, tressauter, se tuer, vibrer, voler, vomir;

colère + 61 verbes:

baver, blanchir, blêmir, *bleuir*, bondir, bouillir, bouillonner, *bourdonner*, brûler, chavirer, chialer, crever, crier, déborder, éclater, écumer, s'empourprer, s'enflammer, *éructer*, s'étrangler, (s')étouffer, s'évanouir, exploser, flamber, frémir, frissonner, *fulminer*, fumer, grimacer, gronder, imploser, marmonner, mourir, mugir, (se) noircir, pâlir, se pâmer, *pester*, piaffer, *piétiner*, pleurer, *se rembrunir,* se ronger, rosir, rougir, rugir, sauter, sautiller, suer, suffoquer, sursauter, tituber, *tonner*, transpirer, trembler, trépigner, tressaillir, vaciller, verdir, vibrer, vociférer, vomir;

tandis que d'autres sont susceptibles d'entrer en collocation avec un nombre de verbes restreint, p.ex.

[57] Le nombre d'occurrences des collocation est aussi variable; ceux qui dépassent, dans notre corpus, 150 occurrences, sont *soulignés*).

***sommeil* + 11 verbes:**

bâiller, *choir*, *crever*, crouler, *dodeliner*, s'écrouler, mourir, succomber, tituber, tomber, se tortiller.

5 ASSOCIATION VINT/N

Leeman 1991 distingue deux groupes principaux de collocations intensives V de N, à savoir celles où l'interprétation du verbe peut être littérale (*rougir de colère*, *crier de peur*) ou doit être non-littérale (*mourir d'ennui*, *rayonner de bonheur*, *brûler d'amour*). L'examen d'un corpus d'exemples numériquement important permet de nuancer cette analyse et de spécifier d'autres mécanismes qui fondent l'association entre le verbe et le nom qu'il intensifie. Ils sont nombreux et variés, ils relèvent aussi bien du domaine cognitif que linguistique; c'est leur nombre et leur diversité qui sont à la base de la multiplication des paradigmes d'association *Vint/N*. Nous passons ici en revue les principaux facteurs qui décident du choix du verbe intensifiant un nom donné.

5.1 Les symptômes stéréotypés (stéréotypie cognitive)

La psychologie des émotions parle des associations systématiques entre les expériences intérieures et leurs manifestations physiologiques typiques. Les symptômes[58] d'un état émotif (p.ex. horripilation, secrétions diverses, changement de couleur de la peau dû à la vasodilatation où vasoconstriction...) témoignent d'un degré d'intensité de celui-ci suffisamment fort pour provoquer une réaction observable de l'organisme. Celle-ci est déclenchée toujours de manière involontaire, mais peut être incontrôlable (*rougir*, *transpirer*...) ou contrôlable, au moins dans une certaine mesure (*trembler*, *vomir*...). Une même émotion peut s'accompagner de tout un spectre de réactions physiologiques (aussi bien en fonction de son intensité qu'en fonction de l'individu qui l'éprouve). Inversement, une même réaction peut apparaître comme symptôme de diverses émotions (*pleurer de joie/de tristesse*). Cette diversité explique l'importance numérique des associations Vint/N fondés sur l'observation de symptômes stéréotypés.

5.2 La stéréotypie culturelle

Un certain nombre de comportements (conscients et contrôlables) correspond à des réactions attribuées, dans une société, à telle expérience intérieure. Elles ont, en général, un fondement physiologique, mais ce sont les conventions sociales qui les consacrent dans le rôle de symptômes culturellement marqués. Ainsi, on peut *rôter de satisfaction*, *siffler* ou *applaudir d'admiration*, *danser* ou *chanter de joie* etc.

5.3 L'hyperbole

L'intensité de la cause est exprimée à travers une manifestation extrême, rarement observable en réalité, qui se réalise seulement dans des situations exceptionnelles. Notre expérience quotidienne dit qu'il est malheureusement possible de *mourir de faim*, que l'on peut, dans de très rares cas, *mourir de peur* ou *de joie* (lorsque l'émotion éprouvée est soudaine et très forte, et l'expérienceur souffre d'une maladie cardiaque), mais que *mourir de honte* ou *d'ennui* n'arrive pas: on a ici un continuum qui va de réactions extrêmes possibles, et parfois réellement constatées, vers l'hyperbole pure et simple.

[58] *Symptôme*: 'conséquence perceptible ou observable liée à un état qu'elle permet de déceler' (cf. la définition du NPR)

5.4 La métaphore

Le langage métaphorique permet de rendre compte des expériences intérieures (domaine-cible de la métaphorisation) en termes d'expériences physiques simples. Les collocations de la série *se gonfler d'orgueil*, *déborder d'enthousiasme*, *exploser de colère*, etc. renvoient à l'image métaphorique du corps humain en tant que contenant, et de l'émotion en tant que substance qui le remplit, et en déborde ou le fait éclater lorsqu'elle atteint un certain degré d'intensité. La métaphore de la lumière est à la base des collocations comme: *rayonner de bonheur*, *pétiller d'énergie*, ou - en dehors du paradigme des noms d'affects - *briller d'esprit* (car la métaphore en question s'applique aussi à l'intensification des noms de qualité). L'être humain est assimilé à une source lumineuse, et le sentiment ou la qualité, à la lumière elle-même, qui a pour caractéristique d'être d'autant mieux observable qu'elle est intense.

5.5 L'isotopie sémantique

Elle se définit par la présence de sèmes récurrents (cf. Rastier 1987) dans divers éléments lexicaux de l'énoncé, comme dans les exemples ci-dessous:

> Vallée du Rhône France: Pour *bourdonner de* plaisir. [...] C'est une petite ruche pleine de bons arômes que Chapoutier a façonné avec des grappes de Syrah fruitées et épicées.
>
> Un seul mot d'ordre, soyez fous de glaces et laissez-vous *fondre de gourmandise*.
>
> *Taureau* (21 avril - 21 mai): Votre Vénus vous mène la vie dure. Prenez-vous par les cornes et affronter [sic] la rupture. On vous a fait vachement trop de vacheries, de quoi meugler de déception.

La présence de ces sèmes peut également être signalée au niveau de la forme, comme dans l'exemple suivant:

> Bouilloire sans fil, capacité 1,5 litres, puissance 2200 watts [...] On adore son témoin de niveau d'eau et le sifflement qui vous avertit que l'eau est chaude. Originale et pratique : elle vous fait déjà bouillir d'envie !!

qui est déjà à la limite d'un jeu de mots (phénomène caractéristique d'ailleurs du discours publicitaire).

5.6 La fonction poétique du langage

Parfois le choix du verbe intensifieur n'est commandé que par le besoin d'être expressif, de marquer le discours de sa personnalité, de jouer avec les mots. L'énonciateur a recours aussi bien aux associations qui apparaissent au niveau sémantique:

> Le mari d'une femme qui venait d'accoucher de jumeaux, s'est mis à bégayer d'émotion.

qu'au niveau formel:

> Et nous rions à crever quand par chance on tombe sur des gens qui trouvent ça tellement louche que ça les fait loucher d'embarras.

qui se réduit parfois à la forme phonétique:

> Des Ramoneurs de menhirs à *hennir de* bonheur [San Antonio]

6 QUELQUES REMARQUES FINALES

Une étude détaillée des collocations recensées lors de notre recherche à partir du Web francophone dépasse le cadre du présent article. Terminons donc par quelques conclusions qui seront développées dans une publication ultérieure (Pilecka 2010a, à paraître).

- La construction *V de N* est une forme d'intensification du prédicat nominal largement attestée en français.

- Le fonctionnement de la structure « prédicat verbal + complément circonstanciel de cause » semble évoluer systématiquement vers l'interprétation où le verbe a la fonction d'intensifieur du prédicat nominal; dans l'usage actuel, cette interprétation est dominante, à quelques exceptions près; nous assistons ainsi à la lexicalisation de la structure et à la grammaticalisation des items verbaux (sens concret, physique → sens abstrait, évaluation de l'intensité).
- Le changement de sens est basé essentiellement sur le mécanisme de la métonymie accompagnée de l'hyperbolisation, et – pour certaines sous-classes de verbes – de la métaphore (cf. Pilecka 2010b, à paraître).
- Le choix de l'intensifieur est, au départ, fondé sur l'expérience cognitive; cependant, il peut être aussi motivé par les facteurs d'ordre essentiellement linguistique. Tous ces mécanismes, qui assurent l'élargissement constant des paradigmes des V intensifieurs et des N intensifiables dans le cadre de la construction *V de N*, sont responsables de l'importance numérique de cette classe des collocations intensives dans l'usage du français contemporain.

7 RÉFÉRENCES

Badyńska-Lipowczan B. (1999). « Analisi semantico-sintattica dei predicati psicologici in francese e in italiano » : *verbes supports*, *opérateurs appropriés* e *classes d'objets*. Thèse de Doctorat, Université de Silésie (non publiée).

Beauchesne J. (2002). *Dictionnaire des cooccurrences.* Montréal : Guérin.

Flaux N. et Van de Velde D. (2000). *Les noms en français, esquisse de classement.* Gap – Paris: Ophrys.

Leeman D. (1991). « *Hurler de rage*, *rayonner de bonheur*: remarques sur une construction en *de* ». *Langue française* 91, p. 80-101.

Mathieu Y. Y. (2001). *Les verbes de sentiment. De l'analyse linguistique au traitement automatique.* Paris: CNRS.

Pilecka E. (2010a à paraître). *Verbes intensifieurs.* Łask : Leksem.

Pilecka E. (2010b à paraître). « Métonymie et métaphore comme moyens d'expression de l'intensité ». Dans *Actes du colloque Des mots et du texte aux conceptions de la description linguistique*, Varsovie, 18-20 juin 2009.

Rastier F. (1987). *Sémantique interprétative*. Paris : PUF.

Wooldridge R. (2005). « Le Web comme corpus d'usages linguistiques ». *Cahiers de Lexicologie* 85. p. 209-225.

Zingle H. et Brobeck-Zingle M.-L. (2003). *Dictionnaire combinatoire du français.* La Maison du Dictionnaire.

LES CORPUS EN QUESTIONS : QUANTITÉ ET QUALITÉ

François Rastier
Directeur de Recherche
CNRS-INaLCO

RÉSUMÉ

L'épistémologie de la linguistique générale et la conception même des langues se trouvent remises en cause par l'essor de la linguistique de corpus.

L'instrumentation permet de nouvelles formes d'objectivation. Elles intéressent notamment de nouveaux observables, comme les normes de discours et de genre, ainsi que les corrélations entre plan du contenu et plan de l'expression qui déterminent la sémiosis textuelle. La méthodologie linguistique se trouve enfin devant le défi d'articuler les méthodes quantitatives et quantitatives.

Mots-clé : objectivation, instrumentation, sémiosis, méthodologie, normes, performances.

Un beau jour du printemps 2004, une journaliste vint me trouver et me demanda de lui parler de l'amour au XXI^e^ siècle. Sur ce sujet éminemment consensuel, le *Journal du CNRS* préparait un dossier interdisciplinaire et l'ouvrage que j'avais dirigé quelques années auparavant, *L'analyse des données textuelles — L'exemple des sentiments dans le roman français (1820-1970)*, avait sans doute semblé me qualifier pour traiter de cette question.

Conscient de mes obligations statutaires, je m'efforçai de répondre, mais je le fis par la question : « Dans quel corpus ? ». Devant le désarroi qui se peignit sur le visage avenant de mon interlocutrice, je me lançai dans des justifications : pour nous, malheureux linguistes, l'amour n'existait que dans les textes et variait avec les discours, les genres et les auteurs. Ainsi n'avait-il rien de commun dans le roman du XIX^e^ siècle, où *amour* trouve pour antonymes *argent* et *mariage*, et dans la poésie de la même époque, d'où l'argent et le mariage restent évidemment absents. Faute d'avoir eu la présence d'esprit de constituer un corpus sur l'amour en ce siècle naissant, je dus enfin confesser mon incompétence. Tout cela dut paraître bien décevant et il n'en résulta qu'un maigre entrefilet dont je suis confus de n'avoir gardé aucun souvenir. Il me parut donc nécessaire d'entreprendre une « action de communication », non plus à propos de l'amour, sujet pourtant porteur, mais de la sémantique de corpus.

1 UNE RUPTURE

De nombreuses collectivités sont de longue date engagées dans une réflexion sur la numérisation et l'analyse assistée des documents : outre bien entendu les sciences de l'information, il faut mentionner entre autres l'histoire, la sociologie, la linguistique, l'archéologie, les études littéraires.

La constitution et l'analyse de corpus est en passe de modifier les pratiques voire les théories en lettres et sciences sociales. Toutes les disciplines ont maintenant affaire à des documents numériques et cela engage pour elles un nouveau rapport à l'empirique. En outre,

la numérisation des textes scientifiques eux-mêmes permet un retour réflexif sur leur élaboration et leurs parcours d'interprétation. Les nouveaux modes d'accès aux documents engagent-ils de nouvelles formes d'élaboration des connaissances ? Les ambitieuses initiatives de numérisation prises au plan national et international peuvent devenir l'occasion et le support d'un projet fédérateur pour les lettres et les sciences sociales.

Le doute positif relève de l'attitude critique nécessaire à toute problématisation scientifique. Il reçoit ici un contenu nouveau, car avec les corpus numériques, les sciences de la culture trouvent de nouvelles perspectives épistémologiques et méthodologiques, alors même qu'elles se trouvent affrontées à des programmes réductionnistes de naturalisation. L'objection récurrente formulée contre leur scientificité tient au caractère non répétable des événements : comme en sociologie, en ethnologie, en psychologie sociale voire en linguistique de l'oral, la présence même de l'enquêteur modifie la situation, on conclut que les sciences de la culture n'auraient donc pas la possibilité d'identifier des causes déterminantes et donc des lois. Or selon le préjugé scientiste qui sous-tend les programmes de naturalisation, la condition nécessaire de la scientificité reste la formulation de lois causales – qu'il faudrait alors chercher dans les substrats physiologiques, neuronaux ou génétiques.

À la classique dualité entre induction et déduction dans les disciplines d'observation, le renouvellement méthodologique favorisé par les corpus numériques engage à substituer le cycle suivant : (i) analyse de la tâche et production des hypothèses ; (ii) constitution d'une archive et sélection d'un corpus de référence ; (ii) élaboration des corpus de travail ; (iii) traitement instrumenté de ces corpus, en contrastant corpus de travail et corpus de référence ; (iv) interprétation des résultats et retour aux sources textuelles pour valider l'interprétation. La puissance propre de ce dispositif heuristique permet de faire émerger de *nouveaux observables* inaccessibles autrement : par exemple, la phonostylistique, jadis condamnée à l'intuition, se voit à présent pourvue de moyens d'investigation par des statistiques sur corpus phonétisés. En outre, l'utilisation d'une instrumentation scientifique (analyseurs, étiqueteurs, etc.) participe du processus d'objectivation : les objets culturels ont beau dépendre de leurs conditions d'élaboration et d'interprétation, les valeurs qu'ils concrétisent peuvent cependant être objectivées comme des faits.

La linguistique de corpus pourvoit ainsi la linguistique d'un domaine où elle peut élaborer des instruments et définir une méthode expérimentale propre : elle ouvre aussi des champs d'application nouveaux et engage un mode spécifique d'articulation entre théorie et pratique. D'une part, alors que la linguistique théoriciste (sans corpus) portait, en extrapolant quelques observations sur des exemples souvent forgés, des jugements universels sur le langage, la linguistique de corpus, sans renoncer à l'élaboration théorique, en limite la portée aux corpus étudiés, et, sans se satisfaire de la seule démarche déductive, procède par essais et erreurs.

En 1999, Noam Chomsky, auteur d'une grammaire universelle, déclarait que la linguistique de corpus n'existait pas, alors même qu'elle était déjà en plein essor : il signalait ainsi qu'elle restait inconcevable pour la linguistique de fauteuil et qu'une rupture épistémologique était consommée. Cette rupture jouit d'une portée générale : en bref, la recherche part d'une diversité constatée, l'unifie dans le point de vue qui préside à la collection du corpus, éprouve enfin son objectivité par l'investigation instrumentée. Ordinairement, la régularité des observables sera portée au crédit du système, la diversité irréductible sera imputée à la contingence du corpus. Toutefois, l'opposition sommaire entre l'unité totalisante et l'irrégularité accidentelle peut sans doute être dépassée dans la description des normes, dont seules les plus générales, parmi l'ensemble des corpus étudiés, seront considérées comme propres à la langue.

2 LES CORPUS ET L'ESPACE DES NORMES

Sans prétendre tirer un bilan prématuré, il semble que la situation nouvelle de la linguistique impose une reconception de la dualité entre linguistique de la langue et linguistique de la parole, qu'il est de tradition d'opposer, tant chez Bally que chez Benveniste, tant en linguistique de l'énonciation qu'en pragmatique, alors que chez Saussure elles sont parfaitement complémentaires.

On a trop souvent réduit les langues à des dictionnaires et des grammaires, voire à des syntaxes. Il faut cependant tenir compte, outre du *système*, des *corpus* (corpus de travail et corpus de référence), de l'*archive* (de la langue historique), enfin des *pratiques* sociales où s'effectuent les activités linguistiques. Pour l'essentiel, une langue repose sur la dualité entre un *système* (condition nécessaire mais non suffisante pour produire et interpréter des textes) et des *corpus* de textes écrits ou oraux[59].

Non contradictoire, la dualité dynamique entre corpus et système constitue la langue dans son histoire. Aussi ne saurait-on assimiler la *langue historique* à la *langue fonctionnelle* (celle qui fonctionne ici et maintenant) en négligeant que la langue historique détermine la langue fonctionnelle dans ses structures et ses contenus. Le corpus de référence sert de médiation entre la langue historique et la langue fonctionnelle, et les textes qui n'appartiennent plus qu'à la langue historique entrent dans l'*archive*. Soit :

Système(s)	*Corpus*
Langue fonctionnelle	Corpus de référence
Langue historique	Archive

Tableau 1 : Instances du système et types de corpus.

En évoquant les corpus et non les signes, nous soulignons que la langue n'est pas un système de signes – comme le serait un code ; Saussure, à qui l'on prête cette définition, ne l'a d'ailleurs jamais formulée. Un signe au demeurant n'a pas de définition intrinsèque : il n'est qu'un *passage*, certes réduit, d'un ou plusieurs textes auxquels il renvoie.

En première approximation, une langue est faite d'un corpus de textes oraux ou écrits et d'un système. Le système reconstitué par les linguistes est une hypothèse rationnelle formulée à partir des régularités observées dans le corpus. Entre le corpus et le système, les normes assurent un rôle de médiation : ancrées dans les pratiques sociales, les normes de discours, de genre et de style témoignent de l'incidence des pratiques sociales sur les textes qui en relèvent[60]. Pour éviter la fausse antinomie entre la langue en tant que système de formes et la langue comme produit d'une culture – qui se traduit dans l'enseignement par la distinction entre « cours de grammaire » et « cours de civilisation » – il paraît préférable de considérer

[59]Dans le corpus d'une langue, les *œuvres* tiennent une place particulière parce qu'elles sont hautement valorisées : par exemple, l'italien est la langue de Dante au sens où son œuvre demeure le parangon historique qui a présidé à la formation de la langue italienne en tant que langue de culture.

[60] Un texte en effet ne peut pas être produit par un système, comme l'a montré l'échec de la grammaire générative appliquée à des systèmes de génération automatique de phrases et *a fortiori* de textes.

que le système comprend des règles et des normes diversement impératives. Par exemple, les règles de la ballade française diffèrent de celles de la ballade anglaise et relèvent du système des normes de la langue française.

Les règles et les normes ne diffèrent sans doute que par leur régime d'évolution diachronique. On sait que les mots (lexies, puis morphèmes) sont issus du figement et de l'érosion de syntagmes ; ce qui vaut pour ces unités linguistiques vaut sans doute pour les règles qui norment leurs relations et les constituent ainsi en unités : les règles sont vraisemblablement des normes discursives invétérées.

En synchronie, toute règle voisine avec des normes qui accompagnent voire conditionnent son application. Ce sont les normes qui permettent et limitent l'application des règles : sans elles, par exemple, on ne pourrait arrêter des enchâssements indéfiniment récursifs mais grammaticalement corrects. On ne peut donc juger de la grammaticalité d'une phrase que si l'on connaît le discours, le genre et le texte où elle est prélevée — outre évidemment la datation et le lieu d'origine de ce texte. Bien qu'élémentaire, cette observation frappe d'inanité les discussions sur l'agrammaticalité et l'asémanticité qui surgissent d'elles-mêmes dès que l'on accepte de discuter de phrases non attestées ou hors contexte.

Ainsi, à la différence de celui d'un langage formel, le système d'une langue est-il en fait pluriel et se décline en régimes structurels différents selon les niveaux et paliers d'analyse. Ses domaines d'organisation locaux ou régionaux ne sont pas unifiés dans une hiérarchie attestant l'existence d'un système unique et homogène, comme en témoigne au demeurant l'évolution continue des langues qui trouvent dans leur hétérogénéité systémique le moteur interne de leur changement perpétuel par perturbations et ajustements.

Non moins plurielles que les instances, les performances se spécifie *a minima* dans la distinction entre corpus (de travail et de référence) et archive[61]. À la grande diversité des pratiques sociales correspond celle des corpus produits en leur sein. Soit, schématiquement :

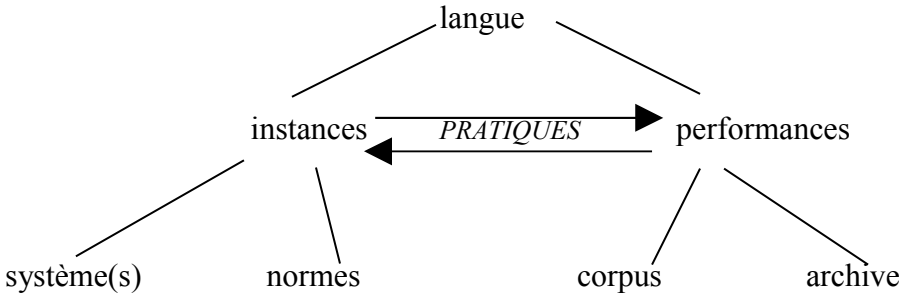


Figure 1 : Instances et performances.

Mise à part la mention de la *langue*, le schéma ci-dessus jouit d'une grande généralité et peut être transposé à des sémiotiques non verbales complexes, comme l'iconologie, par exemple. Nous l'avons d'ailleurs utilisé pour un système d'aide à l'indexation de photographies.

La généralité de ce modèle pourrait lui conférer une portée épistémologique et méthodologique. La dualité *langue-parole* chez Saussure est un cas particulier du rapport entre instances et performances. Au plan méthodologique, la flèche qui va des performances aux instances symbolise l'extraction de régularités ; et la flèche inverse symbolise la caractérisation de singularités, les deux processus restant interdépendants.

[61] Le *corpus de travail* du linguiste n'est qu'une partie du corpus de référence défini par l'ensemble des textes accessibles dans l'empan spatio-temporel considéré. L'ensemble des performances linguistiques non recueillies sur support constitue le *corpus virtuel* de la langue : il garde une incidence, car toute performance modifie peu ou prou les instances normatives qui lui sont associées (règles et/ou normes).

L'objet de la linguistique est la langue, tout à la fois système(s) et corpus, ou plus exactement dualité entre instances et performances. Jusqu'à Chomsky inclus, l'imaginaire grammatical a réduit la langue à son système : c'est d'ailleurs une condition du logicisme traditionnel comme du mécanicisme computationnel. Mais cela suppose que l'être de la langue réside dans la grammaire, voire dans ses structures syntaxiques et non dans ses manifestations empiriques, considérées somme toute comme inessentielles[62].

Définir la langue tant par ses instances que par ses performances, en assumant ainsi la dualité langue-parole, telle qu'elle a été définie par Saussure et oblitérée par les éditeurs du *Cours de linguistique générale*[63], c'est refuser en somme les séparations récurrentes entre l'Être et l'apparence, la puissance et l'acte, le rationnel et le réel, etc. : en bref, sortir de la métaphysique qui a toujours informé la philosophie du langage.

Enfin, au plan épistémologique, il est vraisemblable que la dualité entre instances et performances (ou système(s) et corpus) traduit une dualité de problématiques, l'une de tradition logique et grammaticale, l'autre de tradition rhétorique et herméneutique.

Problématiques	Logico-grammaticale	Rhétorico-herméneutique
Unités privilégiées	Mot, proposition	Texte
Ordres	Règles	Normes
Sémantique	Signification	Sens
Contextualisation	Minimale	Maximale
Instances	Système(s)	Corpus

Tableau 2 : Les problématiques.

La problématique logico-grammaticale privilégie les instances (car elle s'appuie sur une ontologie), alors que la problématique rhétorico-herméneutique privilégie les performances, car elle repose sur une praxéologie. Dans l'histoire des réflexions occidentales sur le langage, tributaires de la problématique logico-grammaticale, les instances dominent les performances : de la théorie scolastique du langage comme faculté qui s'effectue par des actes

[62] La grammaire universelle seule serait selon Chomsky l'objet de la linguistique, et non les langues, ni même le langage : « Pour les linguistes structuralistes et leurs prédécesseurs, l'objet d'étude était le langage et l'analogue le plus proche de la GU [grammaire universelle] était la théorie des propriétés générales de nombreuses (ou de toutes les) langues. Le point de vue que j'ai défendu […] repose sur une attitude tout à fait différente. Le centre d'intérêt est la grammaire. Le langage est une notion dérivée et probablement inintéressante » (1984, p. 21).

[63] Ils ont achevé l'ouvrage en prêtant frauduleusement à Saussure une formule de Franz Bopp (1816) : « La langue en elle-même et pour elle-même ».

(contenus en puissance dans la faculté), on en est par exemple venu à la théorie chomskyenne de la générativité à partir de règles.

La sémantique des textes se propose d'articuler les deux problématiques en reconsidérant la première à la lumière de la seconde, car la première peut être obtenue par restriction drastique de la seconde, alors que la seconde ne peut être obtenue par extension de la première. Plutôt donc que de les considérer isolément comme c'est l'usage, il faut tenir compte du fait qu'elles sont modifiées par leur articulation. Bref, la dualité entre corpus et système(s) n'a rien d'une contradiction : elle est prise dans la dynamique qui constitue la langue dans son histoire et l'institue ainsi en *langue de culture*[64].

3 VERS UN REMEMBREMENT DES DISCIPLINES

En traitant les corpus, la linguistique renoue nécessairement avec les textes, donc avec la philologie et avec l'herméneutique : la philologie pour les établir et les documenter, l'herméneutique pour les interpréter, y compris dans leur dimension intertextuelle.

L'essor de la linguistique de corpus conduit notamment à préciser le rapport entre textes et documents. Alors que la grammaire travaillait sur l'écrit (son nom même l'indique, littéralement), l'oral est une conquête récente de la linguistique ; encore faut-il qu'il soit fixé sur un support, par enregistrement ou transcription, pour devenir l'objet des débats et conjectures propres à l'investigation scientifique. Textes oraux et écrits trouvent leur première unité dans leur statut de documents.

Plus généralement, les différences entre texte et document, bibliothèque et archive, linguistique de corpus et philologie numérique, sont en train de devenir relatives. Le support numérique ne garantit aucune identité à soi : la restitution de l'inscription est sensible aux formats, aux logiciels de visualisation dont les standards évoluent, si bien que la notion philologique d'*herméneutique matérielle* doit ici être comprise indépendamment de tout attendu substantiel.

En perdant son unicité, le document numérique se dépouille des qualités du document unique de l'archiviste : authentifiable, doué par sa continuité matérielle d'une intégrité (même quand il est fragmentaire), non-reproductible, il pouvait faire autorité. À présent, l'affichage par pixels détruit toute continuité matérielle qui empêchait les falsifications. Alors qu'une critique initiale suffisait à établir le document, il faut à présent une critique continue pour maintenir sa fiabilité. L'établissement des significations doit souvent passer par une succession de versions, dont chacune est le support et le résultat d'une opération de lecture. En changeant ainsi de régime, l'objectivation peut progresser sans pouvoir jamais être considérée comme établie, ce qui engage à rompre avec l'objectivisme pour promouvoir une objectivation critique indéfinie.

Toutefois, ce que le document perd en stabilité, il le gagne en biais d'interrogation. Les logiciels appellent une réflexion théorique sur l'étiquetage, sur les rapports entre méthodes qualitatives et quantitatives : on peut par exemple croiser les résultats de plusieurs méthodes pour faire apparaître de *nouveaux observables*. C'est autant aux « gens du texte » qu'aux informaticiens de faire des propositions sur ce point : pour aborder ces questions, la voie technologique et la voie épistémologique n'ont rien de contradictoire.

C'est par la méthodologie comparative que l'on va pouvoir exploiter les possibilités techniques actuelles. Pour fonder cette méthode, lui permettre d'évoluer et lui fixer des objectifs de connaissance, il faut aussi que la linguistique assume sa place parmi les sciences de la culture.

[64] Nous écartons ici les langues purement véhiculaires, comme le *Basic English* : ces artefacts obtenus par restriction drastique de langues de culture restent dépourvus de corpus.

La linguistique au demeurant n'a aucune exclusivité épistémologique dans la réflexion sur les corpus : l'ensemble des sciences sociales et des disciplines littéraires se doivent d'élaborer à leur propos une réflexion coordonnée en gardant leurs objectifs spécifiques. Elles gagnent à des échanges d'expériences, loin d'une interdisciplinarité fusionnelle d'ailleurs illusoire.

Beaucoup cependant reste à faire pour convaincre de la nécessité de travailler sur corpus. La technicité, le détour instrumental, la notion même de méthode expérimentale, inquiètent certains ; l'attachement à la recherche en fauteuil sans sanctions empiriques, parfois même dans des disciplines littéraires la répugnance à l'égard de toute objectivation censée porter atteinte à la subjectivité souveraine des auteurs et des lecteurs, tout cela conduit certains à considérer l'étude des corpus comme un leurre[65].

Ils formulent une objection récurrente : on ne trouve jamais que ce que l'on cherche. Soit ils regrettent par là que l'on vérifie l'intuition sans songer qu'il est parfois difficile de prouver des évidences, ni que cela fait partie de l'ingrate mission des sciences ; soit encore ils estiment qu'on trouve toujours quelque chose : c'est faux, car des résultats bruités peuvent inviter au silence.

De fait, on ne trouve pas toujours ce que l'on cherche, mais souvent autre chose que l'on ne cherchait pas : de nouveaux observables. Certes, on ne trouve trop souvent que ce que l'on sait voir et l'on reste dépendant d'un état de l'art et des problématiques routinières de la « science normale » ; une démarche critique permet cependant de les dépasser ensemble.

Un des problèmes fondamentaux que rencontre la linguistique de corpus est l'interprétabilité des résultats, notamment ceux qu'obtiennent les méthodes quantitatives ; une sémantique de l'interprétation nous semble indispensable pour qualifier les résultats obtenus, car le détour interprétatif est une condition première de l'objectivation.

La première difficulté est de passer des chaînes de caractères à des formes sémantiques, ce qui suppose une méthodologie élaborée, sans quoi l'on en resterait à une lexicométrie limitée. Par exemple, les cooccurrents lexicaux d'un mot-pôle doivent être qualifiés comme des corrélats sémantiques pour pouvoir être considérés comme des lexicalisations partielles d'un thème (cf. l'auteur, éd, 1995). Or, si l'on réussit à construire des formes sémantiques, on peut parvenir en retour à sémantiser les unités de l'expression, même « de bas niveau » comme les ponctèmes ou les phonèmes : on construit ainsi des formes sémiotiques par appariement de formes sémantiques et de formes expressives. Prenons un exemple élémentaire : dans un corpus de romans, le thème de la peur est associé à des points de suspension ; les sèmes /intensité/ et /aspect ponctuel/ sont récurrents dans le contenu de *peur*, ce que marquent des contextes comme *violente*, *brusque*, etc., comme dans le contenu des points de suspension (l'anacoluthe marque une rupture soudaine, d'où le trait /intensité/).

La relation même entre les plans du langage, signifiant et signifié, contenu et expression, reçoit ainsi un éclairage nouveau. Certes, l'analyse morphosyntaxique par étiquetage automatique n'est pas purement " formelle " et s'appuie généralement sur un lexique qui contient des informations sémantiques. Mais cette étape franchie, elle ouvre la possibilité de mettre à jour des corrélations fortes entre les régularités de l'expression et celles du contenu.

[65] « Dans les programmes nationaux pour les SHS [sciences humaines et sociales], l'accent est souvent mis sur les « terrains », les « corpus » et autres « archives ». Certes, on connaît leur importance pour les SHS, mais on peut douter qu'il s'agisse de priorités scientifiques ou sociétales. On n'imagine pas la chimie des matériaux construire un programme de recherche sur les meilleurs gisements de matières premières ou sur les fournisseurs les plus efficaces en poudres ou autres produits de base ». Fontanille, J. (2006) Supplément d'âme ?, *Vie de la recherche scientifique*, 365, pp. 16-17, ici p. 17. Soit, mais les corpus, nécessairement élaborés par leur constitution même, ne sont pas des matières premières ; la chimie des matériaux et la besogneuse alchimie des corpus n'ont d'ailleurs pas le même lustre.

Loin de se limiter aux textes littéraires, la corrélation confirmée entre variables globales comme le discours, le champ générique, le genre et les variables locales (tant morphosyntaxiques que graphiques ou phonologiques), nous a conduit à poser le problème de la *sémiosis textuelle*. On définit ordinairement la sémiosis au palier du signe, et comme un rapport entre signifié et signifiant ; mais on ne s'interroge guère sur les paliers supérieurs, comme si leur sens se déduisait par composition de la signification des signes. Or, un genre définit précisément un rapport normé entre signifiant et signifié au palier textuel : par exemple, dans le genre de la nouvelle, le premier paragraphe est le plus souvent une description, non une introduction comme dans l'article scientifique.

La sémiosis locale et conditionnelle proposée par la langue aux paliers de complexité inférieurs, du morphème à la lexie, ne devient effective que si elle est compatible avec les normes de genre voire de style qui assurent la sémiosis textuelle.

Enfin, l'opposition humboldtienne entre la *forme intérieure* et la *forme extérieure* des textes, qui a fait couler tant d'encre chez les stylisticiens, pourrait recevoir une nouvelle formulation qui la relativise : la forme intérieure, loin d'être un mystère esthétique, est constituée par les régularités jusqu'à présent imperceptibles de la forme extérieure, celle de l'expression, que les moyens théoriques et techniques de la linguistique de corpus permettent à présent de mettre en évidence. En d'autres termes, le contenu n'est pas une mystérieuse représentation mentale : un texte est fait de deux plans, celui des formes sémantiques et celui des formes expressives, dont le genre notamment norme la mise en corrélation. Au sein de chaque plan s'établissent des relations entre forme et fond, de type gestaltiste, qui permettent la perception sémantique et phonologique.

4 LA SÉMIOSIS TEXTUELLE ET LES CORRÉLATIONS ENTRE PLANS DU CONTENU ET DE L'EXPRESSION

Elles sont cruciales pour la sémiotique des textes, car elles permettent d'aborder la question de la sémiosis textuelle. Au plan *graphique*, alors que la ponctuation n'est pas considérée comme sémantique et qu'elle est tout simplement absente de la plupart des grammaires formelles, l'étude en corpus permet de souligner les corrélations entre contenus lexicaux et ponctèmes. Par exemple, dans un corpus romanesque, Evelyne Bourion (2001) a ainsi pu confirmer la corrélation entre des noms de sentiments et les ponctuations dans les contextes où ces noms apparaissent. Ainsi les sentiments ponctuels, brusques, comme la *colère* ou la *joie* sont-ils fortement associés aux points de suspension.

Pour sa part, dans une étude comparative sur Baudelaire, Maupassant, Proust et Duras, Denise Malrieu a ainsi spécifié les contextes de *mer* : « On constate que les ponctèmes après *mer* sont plus fréquents qu'avant *mer*, ces derniers variant de 6% chez Proust à 32% chez Rimbaud, les ponctèmes après variant de 36,2% chez Rimbaud à 53% chez Proust » (2001, p. 11). Il s'agit sans doute, dans la langue littéraire, d'une convention consistant à finir la phrase ou le membre de phrase par un sème /non-borné/, ce qui a un effet d'amplification oratoire. En revanche, des sémèmes comportant le sème /borné/ comme 'mur' n'entraînent aucun résultat comparable.

Ce qui vaut pour les lexèmes vaut aussi pour les grammèmes. Par exemple, dans le corpus littéraire de la société Synapse, le point-virgule et l'imparfait du subjonctif sont associés par une corrélation de 0.44. Cela tient sans doute à leur emploi commun dans les passages d'analyse psychologique : l'imparfait du subjonctif comme le point-virgule sont pour ainsi dire des imperfectifs et peuvent supposer un suspens critique.

Au plan *phonétique*, on constate également des effets de solidarité entre paliers : ainsi, dans les tragédies de Racine, les phonèmes du nom du personnage principal, surtout quand il constitue le titre, sont significativement diffusés sur l'ensemble du texte (cf. Valérie Beaudouin, 2002, 8.3.2). Ainsi, les éléments d'une forme phonique locale se trouvent diffusés pour constituer un fond perceptif global.

En outre, dans son analyse de Racine, Beaudouin (2002, § 8.3.3) a pu montrer que le champ sémantique de la mort était associé à des mètres anapestiques et le champ sémantique de l'amour à des mètres iambiques : la mort est repos, donc les accents sont plus rares, alors que l'amour est passion et se trouve associé à des accents plus fréquents). Mieux encore, le taux d'hémistiches irréguliers selon les actes semble corrélé à la structure narrative globale (cf. Beaudouin, 2002, § 8.3.4).

Les corrélations entre plans du contenu et de l'expression rendent licite la notion de *contextualité hétéroplane* : le contexte d'une unité sur un plan, expression ou contenu, est constitué par d'autres unités sur le même plan, mais aussi sur l'autre. On ressent le besoin d'une théorie qui puisse penser ces corrélations, c'est-à-dire d'une linguistique informée par une sémiotique textuelle.

L'étude des textes littéraires est ici particulièrement révélatrice, car ils multiplient les rapports entre global et local, par des structures en abîme, notamment, tout comme les rapports entre plans du contenu et de l'expression : cela est évident en poésie, par exemple dans l'usage de la rime. Ainsi, par leur complexité, ils « signifient » plus, ce qui leur vaut d'être relus.

Les mêmes types de corrélations sont toutefois à l'œuvre dans des corpus non littéraires. Ainsi le projet européen *Princip.net* de détection automatique de sites racistes (cf. l'auteur, 2006) a mis à profit des critères de « bas niveau » comme la ponctuation (un antiraciste ne redouble jamais un point d'exclamation), la casse (un antiraciste n'écrit jamais une phrase en majuscules), les polices de caractères, voire les codes html (les images sont caractéristiques des sites racistes).

Les corrélations entre plans du contenu et de l'expression ont aussi un enjeu immédiat pour les applications comme la catégorisation de documents, la détection automatique de sites, etc. En pratique, elles permettent, dès lors que la catégorisation des documents du corpus d'apprentissage tient compte d'une classification évoluée, d'éviter des traitements sémantiques complexes et aléatoires.

Les méthodes de recueil et d'analyse de corpus ainsi mises au point s'adaptent aussi à d'autres sémiotiques, où le repérage de corrélations entre contenu et expression permet par exemple l'identification rapide des genres. Ainsi dans une application d'assistance à l'indexation d'images fixes (projet Semindex, ENSTb-Télécom), les genres d'un corpus de presse *people* ont pu être discriminés en fonction de critères d'expression : par exemple, toute photo à gros grain peut être classée dans le genre de l'*indiscrétion* — qui suppose l'usage d'un téléobjectif.

5 QUANTITÉ ET QUALITÉ EN SÉMANTIQUE DE CORPUS

Venons-en pour finir au problème méthodologique : comment articuler traitements quantitatifs et descriptions qualitatives ?

Mesure pour mesure. — La linguistique de corpus est s'appuie sur des méthodes quantitatives (pour l'essentiel statistiques). L'évolution, en France, de la lexicométrie vers la textométrie montre que l'on veut passer d'une « mesure des mots » à une « mesure des textes ». Mais comment concevoir cette mesure sans quelques précautions ?

(i) Ce qui est mesurable n'est pas forcément intéressant – bien que la quantification ait fini par devenir synonyme de respectabilité scientifique (les premières statistiques sociologiques, par Durkheim, ont marqué une rupture dans la tradition « littéraire » des sciences de la culture).
(ii) Ce qui est fréquent ne l'est pas forcément non plus ; notamment, les fréquences absolues sont inutilisables. Dans un texte, les mots les plus fréquents sont des grammèmes qu'on retrouve dans tous les autres textes de la langue. Les lexèmes les plus fréquents peuvent (parfois) permettre d'identifier le domaine sémantique, mais pas les traits caractérisants du texte. Ils constituent en somme le « fond sémantique », alors que les tâches intéressantes consistent à identifier les formes sur lesquelles elles se détachent.
(iii) Les traits « de forme » n'ont pas de poids statistique déterminable, du moins dans l'état de l'art.
(iv) Les unités rares, de fréquence 1 (hapax), ou absentes sont tout aussi intéressantes et souvent caractérisantes (le mot *ptyx*, présent uniquement chez Mallarmé, a fini par symboliser son œuvre ; le mot *homme* reste absent des sites racistes, cf. l'auteur, 2006).
(v) Les éléments les plus caractérisants sont des corrélations qui peuvent relier des éléments peu fréquents. Les événements linguistiques (si par exemple l'on cherche à détecter les créations de concepts), sont des syntagmes inédits, donc des « hapax » combinatoires – la néologie n'ayant qu'un rôle secondaire.
(vi) Le qualitatif peut échapper à tout dénombrement : si les chaînes de caractères peuvent être dénombrées, il n'est pas certain que les unités sémantiques soient assez discrètes et stables pour l'être de la même manière. De simples cooccurrences sans poids statistique peuvent être révélatrices (ex. *mari* et *amant*, *amour* et *argent*, cf. l'auteur, 2004).

Qualifier les nouveaux observables. — La force heuristique de la linguistique de corpus tient à ce qu'elle peut mettre en évidence voire « faire émerger » de nouveaux observables, notamment : (i) Des associations entre éléments qualitatif et quantitatif (ex. en corpus littéraire, les hapax sont associés aux pronoms de troisième personne). (ii) Des inégalités qualitatives, notamment dans la topographie textuelle (ex. les rafales ; les inégalités distributionnelles dans les sections découpées en déciles). (iii) Des associations entre unités sémantiques (thèmes) et unités expressives (ponctuation), même « de bas niveau » (casse).

Détaillons l'exemple du lien entre hapax et pronoms, décelé par Étienne Brunet : « Les textes qui font appel à la première personne, et plus encore ceux qui s'adressent à la deuxième personne, répugnent à l'emploi des hapax, lesquels par contre abondent dans l'entourage de la troisième personne. »[66]. Cette constatation factuelle est troublante : d'une part, aucun linguiste n'a jamais songé à lier l'usage d'une personne pronominale avec la rareté du vocabulaire qui l'entoure ; d'autre part, une personne pronominale est un phénomène qualitatif, un hapax est un phénomène quantitatif, d'ailleurs relatif au texte : un mot très rare peut être employé à plusieurs reprises dans un texte, un mot généralement fréquent peut avoir le statut d'un hapax dans un autre, dès lors qu'il n'apparaît qu'une fois. Ébauchons une hypothèse : si l'étude de Brunet porte sur le vocabulaire français, le corpus sur lequel il s'appuie, celui de Frantext, est pour l'essentiel constitué d'œuvres littéraires, notamment de romans. Or, quand deux personnages de roman parlent, leur vocabulaire est une image du français parlé qui, pour être littéraire, n'en est pas moins simplifiée par rapport aux passages non dialogués. Le *tu* fait dont fuir les hapax. En revanche, le *je* peut se trouver dans des passages de méditations, de descriptions, de remémorations où le narrateur peut employer un vocabulaire plus détaillé sans être taxé de pédantisme. *A fortiori*, la troisième personne se trouve en général déliée de

[66] . Étienne Brunet, *Le vocabulaire français de 1789 à nos jours (I)*, Paris –Genève, Champion – Slatkine, 1981, p. 75.

toute intériorité et affranchie des mots simples qui s'emploient dans la conversation ou dans le monologue intérieur.

En formulant cette hypothèse, je souhaite illustrer que les corrélations constatées doivent pouvoir être interprétées pour accéder au rang de faits nouveaux. En outre, elles ne sont pas des « faits de langue » bruts, mais doivent être d'abord rapportées aux genres et aux discours présents dans le corpus. Retenons en somme qu'un nouvel observable est un générateur d'hypothèses et éventuellement un précieux destructeur d'évidences.

Dépasser la contradiction quantité/qualité. — Pas plus que le fréquent n'est quantitatif, le rare ne se confond avec le qualitatif. Il n'y a pas d'opposition entre quantitatif (positiviste) et qualitatif (élitiste), mais une complémentarité : ainsi, le résultat quantitatif peut confirmer l'hypothèse qualitative. Fréquente ou rare, toute donnée numérique, fût-elle un zéro, doit être rapportée à une donnée textuelle. Or une donnée textuelle, c'est ce qu'on se donne. Par construction, elle est le lieu d'interaction de quatre pôles : Contenu et Expression, Point de vue et Garantie (cf. l'auteur, 2008). Ordinairement, à partir d'une expression, on doit reconstituer et qualifier les trois autres pôles pour objectiver la donnée. Les « données textuelles » sont ainsi le résultat d'une interprétation. D'autant plus que les sorties logicielles sont souvent ininterprétables : par exemple le résultat graphique d'une analyse factorielle n'obéit à aucune métrique simple, et il faut bien connaître le corpus pour pouvoir l'interpréter. Ce n'est donc pas l'instrumentation qui permet l'interprétation, mais l'inverse.

Pour une sémantique instrumentée. — Les besoins méthodologiques sont d'autant plus grand que les sémantiques dont on dispose sont pour l'essentiel (i) lexicales, (ii) à postulats mentalistes (sémantique cognitive, théorie du prototype, etc.), (iii) sans protocoles expérimentaux. La linguistique de corpus peut cependant faire avancer la réflexion sémantique au niveau méthodologique (pratique) comme au niveau épistémologique (théorique).

a) Le sens étant fait de différences, le détour méthodologique par l'instrumentation permet de construire des différences : entre mots, entre passages, entre textes, entre auteurs, genres, discours.

La pertinence n'émerge pas du quantitatif, mais de la rencontre entre deux horizons : la pertinence « subjective » déterminée par la tâche et la pertinence « objective » propres aux inégalités qualitatives au sein des textes et entre les textes.

b) Au niveau épistémologique, le détour expérimental permet l'objectivation : (i) en infirmant ou en confirmant des hypothèses, (ii) en faisant ressortir les régularités structurelles de l'objet, quand diverses procédures instrumentales parviennent à des résultats concordants malgré les différences de matériau expérimental, d'échelle, etc.

6 RÉFÉRENCES

Beaudouin V. (2002). *Rythme et mètre du vers classique. Corneille et Racine*. Paris : Champion.

Chomsky N. (1984). La connaissance du langage, *Communications*, 40.

Malrieu D. (2001) *Analyse sémantique des contextes d'un lexème avec aides logicielles : l'exemple de* la mer *dans des textes littéraires.* 21 p.
infolang.u-paris10.fr/modyco/textes/**malrieu**/DM_Contextes_**Mer**.pdf

Rastier F. (éd.) (1995). *L'analyse thématique des données textuelles — L'exemple des sentiments*. Paris : Didier.

Rastier F. (2004a). « Doxa et lexique en corpus - Pour une sémantique des « idéologies » ». *Actes des Journées scientifiques en linguistique 2002-2003 du CIRLLEP*. Reims : Presses Universitaires de Reims.

Rastier F. (2004b). « Ontologie(s) ». *Revue de l'Intelligence Artificielle*, Numéro spécial Informatique et terminologies, 18, p. 16-39.

Rastier F. (2005). « Enjeux épistémologiques de la linguistique de corpus ». Dans G. Williams (éd.), *La Linguistique de corpus*. Rennes : Presses Universitaires de Rennes. p. 31-46.
Rastier F. (2006). « Sémiotique des sites racistes ». *Mots*, 80, p. 73-85.
Rastier F. (2007). « Passages ». *Corpus*, 6, p. 127-162.
Rastier F. (2008a). « Que cachent les données textuelles ? ». Conférence invitée, *Actes des IXe JADT*, Presses Universitaires de Lyon, édités par Serge Heiden et Bénédicte Pincemin, tome I, p. 13-26.
Rastier F. (2008b). « Sémantique du Web vs Web sémantique ». *Syntaxe et sémantique*, 9, p. 15-36.
Saussure F. de (1972). *Cours de linguistique générale.* Paris : Payot.
Saussure F. de (2002). *Écrits de linguistique générale.* Paris : Gallimard (éd. Simon Bouquet et Rudolf Engler).
Valette M. (2008). « Pour une science des textes instrumentée ». *Syntaxe et sémantique*, 9, introduction, p. 9-14.

PROPOSITIONS POUR L'ENRICHISSEMENT SÉMANTIQUE DE CORPUS TEXTUELS

Coralie Reutenauer, Mick Grzesitchak, Evelyne Jacquey et Mathieu Valette
ATILF – CNRS, Nancy Université (UMR 7118)

ABSTRACT

The study implements a process of corpus annotation with semes relying on a textual semantics background. The incentive is to validate this process and also to analyze the additional information coming from this semantic annotation.

RÉSUMÉ

La présente étude met en œuvre une procédure d'annotation de corpus en traits sémantiques inspirée de principes de la sémantique textuelle. Elle cherche à évaluer d'une part la validité de l'annotation, d'autre part ses apports par rapport à une approche lexicale classique à partir d'un outil lexicométrique classique, le calcul des spécificités.

1 CONTEXTE ET OBJECTIFS

Le débat sur les métadonnées et l'enrichissement des corpus est soutenu. Tandis que la tradition française de la textométrie a longtemps considéré la forme comme unité de référence (cf. Brunet 2000, Mellet 2002 pour une discussion), le Traitement Automatique du Langage tend à lemmatiser systématiquement les corpus. Avec l'amélioration des techniques informatiques, on assiste à l'émergence de corpus multi-annotés et d'outils capables de traiter différents niveaux d'analyse, principalement morphosyntaxiques tels que les lemmes, les parties du discours et les catégories syntagmatiques (cf. par exemple le CorpusReader de Loiseau 2005). Si les outils d'annotation morphosyntaxique ont atteint une certaine maturité (Habert 2005), l'annotation sémantique reste peu dotée. Certes, le TAL et la Recherche d'Information confient parfois aux ontologies le soin de rendre compte de ce niveau, mais leur statut linguistique est très contesté (Slodzian 1999) – le peu d'attention qu'attirent ces ressources dans la communauté des statistiques textuelles et de la textométrie est sans doute l'indice de leur inadéquation. Récemment, une approche fondée sur l'exploitation de sèmes en guise de traits sémantiques a vu le jour. Un dictionnaire de sèmes qui se démarque de l'approche ontologique a été réalisé à partir d'une extraction depuis le *Trésor de la Langue Française informatisé* (Pierrel et Dendien 2003) (Valette *et al.* 2006, Grzesitchak *et al.* 2007, Valette 2008). Inspirés de la sémantique textuelle (Rastier 2001), les présupposés théoriques qui ont motivé la réalisation de cette ressource relèvent de conceptions partagées par la textométrie comme, par exemple, le primat accordé à la cooccurrence sur l'unité isolée (Mayaffre 2008).

L'objectif de cet article est d'évaluer, dans ce contexte, l'apport de l'annotation sémique pour la linguistique de corpus, en particulier pour la textométrie. L'expérience menée s'appuie sur une analyse lexicométrique classique et répandue, le calcul de spécificités. Elle repose sur

la confrontation d'un corpus de formes (sans annotation sémantique) au même corpus enrichi en traits sémantiques.

2 CORPUS : DU LEXICAL AU SÉMIQUE

2.1 Présentation du corpus

Le corpus utilisé est issu du discours journalistique. Il est constitué de 1587 articles de presse, tirés de deux quotidiens nationaux aux lignes éditoriales très contrastées, *Le Figaro* et *l'Humanité*. Les articles sélectionnés ont pour sujet la crise économique et financière ; ils couvrent la période de septembre 2008 à février 2009.

Le corpus se présente sous forme de deux versions parallèles : la version lexicale, d'un million d'occurrences de formes, et la version sémique (cf 2.2), de 23 millions d'occurrences de ce que nous qualifierons, en l'absence de validation systématique par le sémanticien, de "candidats-sèmes" par analogie aux *candidats-termes* de la terminologie. La taille du vocabulaire est du même ordre de grandeur dans les deux versions. Les informations principales sur la taille des deux versions du corpus sont récapitulées en figure 1.

	Total	*le Figaro*	*l'Humanité*
Nombre d'articles	1 587	928	659
Formes (version lexicale du corpus)			
Nombre d'occurrences	920 551	533 117	387 434
Nombre de formes	35 147	26 433	23 203
Candidats-sèmes (version sémique du corpus)			
Nombre d'occurrences	23 198 346	13 329 284	9 869 062
Nombre de candidats-sèmes	29 661	25 741	24 434

Figure 1 : Informations sur la taille du corpus

2.2 Constitution d'une version sémique du corpus

La constitution d'une version sémique du corpus est réalisée à partir d'une procédure mise au point par (Grzesitchak *et al*., 2007). Le schéma de la figure 2 récapitule les différentes étapes. Le corpus initial est étiqueté en morpho-syntaxe, lemmatisé et les mots-outils y sont éliminés. L'entrée correspondant à chaque lemme est recherchée dans une ressource lexicographique, le *Trésor de la Langue Française informatisé* (TLFi, Dendien & Pierrel, 2003). Seuls les substantifs, verbes, adjectifs et adverbes des définitions sont conservés. Chaque élément extrait de la définition est considéré comme un candidat-sème ; l'ensemble des candidats-sèmes issus d'une entrée constitue le sémème du lemme d'origine. Ce sémème est substitué au lemme en question dans le corpus. Ainsi, par substitution lemme par lemme, on obtient la version sémique du corpus.

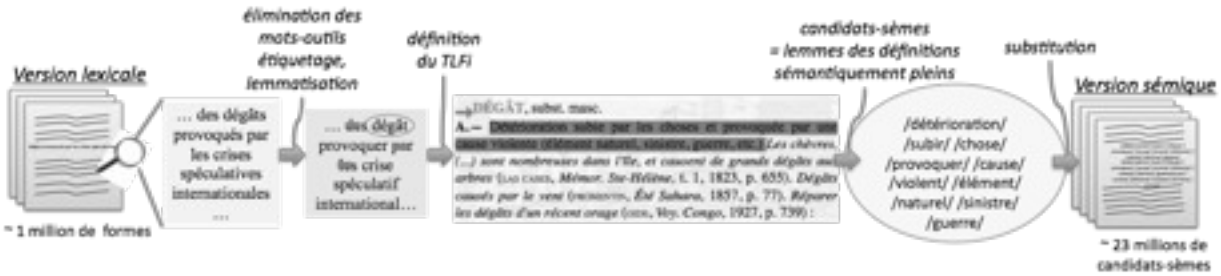


Figure 2 : Schéma de génération de la version sémique du corpus

3 ANGLES ET OUTILS D'APPROCHE DU CORPUS ANNOTÉ

Différents axes d'observation du corpus ont été retenus. Le développement de ces axes repose sur l'identification de contenu sémantique saillant, à l'aide du calcul des spécificités.

3.1 Outil mathématique : le calcul des spécificités

Le calcul des spécificités, décrit dans (Lafon, 1984), a pour but de déterminer le degré de surreprésentation ou de sous-représentation statistique d'une unité dans un sous-corpus par rapport à la totalité du corpus. Issu du modèle hypergéométrique, ce calcul utilise des comparaisons entre partie et tout. Pour une unité et un sous-corpus donnés, il nécessite les informations suivantes : le nombre d'occurrences de l'unité dans le sous-corpus ; le nombre d'occurrences de l'unité dans l'ensemble du corpus : la taille du sous-corpus ; la taille du corpus. Si l'unité est surreprésentée dans le sous-corpus (nombre d'occurrences dans le sous-corpus supérieur à la valeur modale), la valeur de la spécificité est calculée à partir de la probabilité d'avoir au moins le nombre d'occurrences observé ; cette spécificité est positive. Si l'unité est sous-représentée, la valeur est calculée à partir de la probabilité d'avoir au plus le nombre d'occurrences observé ; cette spécificité est négative. Les valeurs des spécificités sont des entiers.

Dans cette étude, le calcul des spécificités est implémenté par le logiciel de textométrie Lexico3 (Salem *et al.,* 2003). Les valeurs sont calculées au-delà d'un seuil de fréquence, fixé ici à 10. Toute unité (candidat-sème sur le plan sémique, forme sur le plan lexical) se voit affecter une spécificité si elle respecte les conditions de seuil.

3.2 Application au corpus

Le calcul des spécificités intervient dans deux approches, une approche globale et une approche locale. L'approche globale se situe à l'échelle d'un journal dans son ensemble. Elle vise l'étude de l'influence des lignes éditoriales de chaque journal. Elle est réalisée dans une perspective de validation de l'annotation sémique.

L'approche locale se focalise sur les paragraphes contenant un syntagme déterminé. Elle cherche à faire émerger des éléments issus du voisinage de l'unité ciblée et susceptibles de caractériser celle-ci. L'unité choisie ici est le syntagme *économie réelle*. Il est présent 176 fois dans 168 paragraphes. La taille du voisinage de cooccurrence est le paragraphe.

Les deux approches reposent d'une part sur une confrontation du plan sémique à une référence issue d'une évaluation intuitive, d'autre part sur une confrontation du plan sémique au plan lexical à l'aide des spécificités. Le calcul des spécificités est donc appliqué à la fois sur le plan sémique et sur le plan lexical.

Dans l'approche globale, le corpus est partitionné en deux selon la source, *L'Humanité* et *Le Figaro*. Ces deux sous-corpus servent tour à tour de référence pour le calcul de spécificité.

Notons que, par complémentarité des deux sous-corpus, une valeur positive sur une unité donnée dans un corpus correspond à la valeur opposée dans l'autre sous-corpus.

Dans l'approche locale, le sous-corpus de référence est, sur le plan lexical, l'ensemble des paragraphes contenant le syntagme *économie réelle,* et, sur le plan sémique, ce même ensemble de paragraphes converti en candidats-sèmes par la procédure d'annotation.

4 APPROCHE GLOBALE ET VALIDATION DE L'ANNOTATION SÉMIQUE

Les résultats obtenus dans l'approche globale, c'est-à-dire à l'échelle d'un journal, se présentent sous forme de listes de spécificités à la fois vastes et diversifiées, avec plus de 2000 formes sur le plan lexical et plus de 7000 candidats-sèmes pour un seuil de spécificité de 2. Deux approches manuelles ont été mises en place pour exploiter ces listes : l'observation des unités les plus spécifiques et un filtrage par catégories déterminées à la lecture.

4.1 Observation des unités les plus spécifiques

Le choix d'un seuil de spécificité élevé, de 20 sur le plan lexical et de 30 sur le plan sémique permet de réduire la liste considérée respectivement à quelques dizaines de formes lexicales et à une centaine de candidats-sèmes environ.. Les résultats présentés en figures (3a) et (3b) correspondent aux formes et candidats-sèmes les plus spécifiques de *l'Humanité.*

Forme	Spécificité	Forme	Spécificité	Forme	Spécificité
travail	≥50	syndicats	30	direction	22
CGT	≥50	communiste	29	des	22
salariés	≥50	sociales	28	Parti	21
salaires	≥50	social	27	capitalistes	21
PCF	48	on	27	pouvoir	21
capitalisme	43	emploi	26	public	21
gauche	41	politiques	25	dividendes	20
sociale	41	pôle	24	les	20
communistes	31	traité	23		
et	31	travailleurs	23		

Figure 3a : Formes lexicales les plus spécifiques de l'Humanité

Forme	Spf	Forme	Spf	Forme	Spf	Forme	Spf
organiser#v	≥50	engendrer#v	≥50	correctionnel#adj	40	syndiquer#v	34
propriété#subst	≥50	réorganisation#subst	≥50	échoir#v	40	formation#subst	34
prérogative#subst	≥50	syndic#subst	≥50	doctrine#subst	39	outil#subst	34
emploi#subst	≥50	communisme#subst	≥50	profit#subst	39	signer#v	33
employé#subst	≥50	rétribuer#v	≥50	utilisable#adj	39	gros#adj	33
commun#adj	≥50	indivision#subst	≥50	capitaliste#subst	39	égalité#subst	33
D=sociopolitique	≥50	critère#subst	≥50	concerner#v	39	catégoriel#adj	33
laborieux#adj	≥50	partageux#adj	≥50	condition#subst	39	entretenir#v	32
solidaire#adj	≥50	qualification#subst	50	gauche#adj	39	définir#subst	32
colonie#subst	≥50	voisinage#subst	50	inconfortable#adj	39	pension#subst	32
vénal#adj	≥50	communiste#adj	48	obligatoire#adj	38	peuplement#subst	32
employeur#subst	≥50	ouvrier#adj	48	venu#adj	38	dépendre#v	32
issir#v	≥50	communautaire#adj	47	militer#v	38	connexe#adj	32
culturel#adj	≥50	subvenir#v	47	paroisse#subst	37	D=droit	31
travailleur#subst	≥50	baser#v	46	voûter#v	37	besoin#subst	31
rémunération#subst	≥50	prôner#v	45	production#subst	37	tâcher#v	31
société#subst	≥50	pansage#subst	45	ligue#subst	37	entreprise#subst	31
dense#adj	≥50	ferrage#subst	45	recherche#subst	37	volontaire#adj	31
adaptation#subst	≥50	boeuf#subst	44	favoriser#v	37	profitable#adj	31
mondain#adj	≥50	expulsion#subst	43	affectation#subst	36	social#subst	31
strict#adj	≥50	but#subst	43	faculté#subst	36	régional#adj	31
mutation#subst	≥50	exercice#subst	42	bénéficier#v	35	protection#subst	30
amélioration#subst	≥50	patronal#adj	42	régulier#adj	35	notice#subst	30
salarier#v	≥50	besogne#subst	42	compromission#subst	35	honnête#adj	30
social#adj	≥50	remise#adj	41	subsistance#subst	35	notoire#adj	30
défavorisé#adj	≥50	utérus#subst	41	syndicat#subst	34	foetus#subst	30
communauté#subst	≥50	dilatation#subst	41	exercer#v	34	moyen#subst	30
classe#subst	≥50	col#subst	41	prisonnier#adj	34	richesse#subst	30
salarié#subst	≥50	capitalisme#subst	40	affirmé#adj	34		
marx#np	≥50	matériel#adj	40	sensé#adj	34		

Figure 3b : Candidats-sèmes les plus spécifiques de l'Humanité

Parmi les unités les plus spécifiques de *L'Humanité,* les orientations sociopolitiques du journal émergent nettement, pour les formes lexicales comme pour les candidats-sèmes : les problématiques des classes sociales, de la gauche, militantisme, syndicalisme et champ sémantique du travail et de l'emploi, sont très présentes. Par ailleurs, un certain nombre de candidats-sèmes renvoient à des notions moins classiques, plus latentes. C'est par exemple le cas de /prérogative#subst/ ou /vénal#adj/, de spécificité supérieure à 50, et dont l'équivalent est absent au niveau des formes les plus spécifiques. Cet enrichissement sur le plan sémique n'est néanmoins pas sans contrepartie : le bruit augmente au niveau des candidats-sèmes. Il provient de diverses sources :

57. de l'absence de filtrage domanial, à l'origine de candidats-sèmes non pertinents. Citons par exemple le cas d' /utérus#subst/, provenant de la définition de la forme lexicale *travail* rattachée au domaine de l'obstétrique (travail lors de l'accouchement). Ces traits non pertinents soulèvent la question du filtrage lors de l'annotation sémique, aussi bien domanial qu'interne à une définition.
58. de candidats-sèmes provenant du métalangage lexicographique, comme /concerner#v/
59. de candidats-sèmes non interprétables, par exemple en raison de leur caractère prédicatif (par exemple /favoriser#v/)

Les résultats sur les unités les plus spécifiques du *Figaro* indiquent également des contenus sémantiques en adéquation avec la ligne éditoriale du quotidien, avec un focus marqué sur les marchés et un regard tourné vers les puissances capitalistes. Le bruit est cependant plus important dans *Le Figaro* que dans *L'Humanité*.

Ainsi, l'étude des unités les plus spécifiques fait émerger des contenus sémantiques caractéristiques des deux journaux aussi bien sur le plan lexical que sémique, en conformité avec l'évaluation intuitive élaborée à partir de la lecture des articles. Cependant, la présence de bruit, accru lorsque la spécificité diminue, invite à raffiner l'approche de la liste de spécificités. La seconde approche, qui propose d'observer l'information lexicale et sémique au prisme de catégories définies manuellement, se situe dans cette optique.

4.2 Répartition en catégories

Lors de la constitution du corpus, le parcours des articles a permis de dégager des valeurs sémantiques caractéristiques de l'un ou l'autre des deux journaux. Ceux-ci ont servi à agencer les unités lexicales ou sémiques en catégories. Pour chaque catégorie, des formes et des candidats-sèmes pertinents sont sélectionnés. Cette sélection repose également sur un critère de rapprochement facile des formes et des candidats-sèmes. Par exemple, la sélection des formes lexicales *travailleur* et *travailleurs* fait pendant à celle de /travailleur#subst/, /travailleur#adj/ sur le plan sémique.

L'objectif est d'observer la convergence des formes et des candidats-sèmes sur des axes sémantiques majeurs. Notons que la sélection effectuée n'est pas exhaustive, il n'est donc pas question d'étudier l'expansion des unités du plan lexical vers le plan sémique, ni quantitativement, ni qualitativement.

Les résultats sont structurés en quatre grandes catégories : acteurs ; dimension nationale et internationale ; vocabulaire économique ; travail et activité. Ces grandes catégories sont subdivisées en sous-catégories. Une partie des résultats, extraite de la catégorie "acteurs", est présentée en figure (4). A chaque catégorie est associée une liste de candidats-sèmes ou de formes affectés de leur spécificité. Pour chaque journal, seules les spécificités positives sont indiquées. Par complémentarité, une spécificité positive pour un journal correspond à son opposé pour l'autre journal. Ainsi, la forme s*yndicat* est de spécificité +6 pour l'Humanité : elle sera donc de -6 pour le Figaro, et le coefficient 6 est reporté dans la colonne correspondant à *L'Humanité*.

Forme lexicale	Spécificité pour *Le Figaro*	Spécificité pour *L'Humanité*	Candidat-sèmes	Spécificité pour *Le Figaro*	Spécificité pour *L'Humanité*
Catégorie ACTEURS					
Syndicats					
syndicat		6	syndic#subst		≥50
syndicats		30	syndicat#subst		34
syndical		11	syndicalisme#subst		22
syndicale		12	syndical#adj		13
syndicales		12	intersyndical#adj		3
syndicalisme		4			
syndicaliste		12	militant#subst		4
syndicalistes		9	militer#v		38
syndicaux		6	militant#adj		25
intersyndicale		3			
Thibault		9	thibault#nam		9
délégué		10	délégué#subst		6
délégués		3			
CGT		≥50			
CFDT		9			
CFE-CGC		3			
CGC		4			
CFTC		6			
FO		6			
Partis					
PCF		48	parti#subst		11
PS		3			
UMP		5	ump#nam		6
Parti		21			
Acteurs socio-économiques et catégories socio-professionnelles					
agriculteurs		3	D=agriculture[1]		2
paysans		3	agriculteur#subst		6
ouvriers		5	ouvrier#adj		48
ouvrière		4	ouvrier#subst	12	
travailleur		4	travailleur#subst		≥50
travailleurs		23	travailleur#adj		8
salarié		7	salarié#subst		≥50
salariés		≥50	salarier#v		≥50
			salarié#adj		10
patron	7		patronat#subst		9
patrons		3			
patronat		10	patronal#adj		42

[1] D=... sert à désigner un domaine

Figure 4 : Spécificités des unités de la catégorie "acteur"

Les résultats obtenus indiquent d'une part une adéquation entre les observations humaines et les tendances indiquées par les spécificités, d'autre part une convergence entre plan sémique et lexical. Par exemple, la notion de syndicat apparaît comme très spécifique de *l'Humanité* aussi bien à travers les formes qu'à travers les candidats-sèmes. Ainsi, la convergence entre évaluation manuelle, plan lexical et plan sémique au niveau de grandes tendances valide l'annotation sémique. L'existence de différences plus fines au sein des catégories souligne un apport propre de l'annotation sémique, dont l'étude plus détaillée fait l'objet de l'approche locale.

5 APPROCHE LOCALE ET APPORTS DE L'ANNOTATION SÉMIQUE

Nous avons cherché à confronter le sens d'un mot-pôle, *économie réelle*, tel qu'il se dégage à la lecture à celui qui émerge d'une part à travers ses cooccurrents lexicaux et d'autre part à

partir d'un faisceau d'unités de sens issues du voisinage sémique. A la lecture des paragraphes, la crise économique apparaît comme une pathologie contagieuse ou comme une catastrophe naturelle se propageant de la sphère financière, considérée comme virtuelle, à la sphère industrielle, correspondant à l'économie dite réelle. Ces observations du lecteur ont servi par la suite à guider et à valider les analyses. Celles-ci portent dans un premier temps sur les unités les plus spécifiques du voisinage d'*économie réelle* et dans un second temps sur le voisinage filtré par des catégories déterminées à la lecture.

5.1 Unités les plus spécifiques du voisinage d'économie réelle

L'observation des listes de formes et candidats-sèmes les plus spécifiques du voisinage *d'économie réelle,* disponibles en figures (5a) et (5b), fait ressortir nettement une dimension économique et financière (présence par exemple des candidats-sèmes /budget/, /argent/, /capitaliste/, /économie/ sur le plan sémique, et des formes *financière, financier, profit* sur le plan lexical). De même, la sphère réelle apparaît à travers les unités les plus spécifiques, surtout sur le plan sémique, à travers des candidats-sèmes comme /chômage/, /bien/, /ressource/, /surproduction/. La notion de choc est également présente (/collision/, /répercussion/, /effondrement/ sur le plan sémique ; *impact* sur le plan lexical), de même que celle de propagation ou même de maladie (forme *contagion* ; candidats-sèmes /contagion/, /dysfonctionnement/ et /pathologique/). Les idées sensibles à la lecture se retrouvent ainsi au niveau des unités les plus spécifiques, sur le plan lexical et de façon encore plus marquée sur le plan sémique. Cependant, le nombre de formes ou candidats-sèmes associés à une idée donnée reste relativement limité, du fait de la taille volontairement réduite de la liste d'unités les plus spécifiques, d'où la mise en place d'une seconde approche des cooccurrents lexicaux et sémiques d'*économie réelle*. Cette seconde approche vise à établir des catégories partant d'idées dégagées de la lecture ou partant de l'observation des unités les plus spécifiques, à affecter des unités à ces catégories puis à confronter l'ensemble des représentants lexicaux et sémiques d'une même catégorie.

Forme (1/3)	**Spf**	**Forme (2/3)**	**Spf**	**Forme (3/3)**	**Spf**
l	20	financier	7	salaires	6
financière	15	conséquences	7	effets	6
impact	11	profits	7	revenus	6
crise	9	richesses	7	contagion	6
récession	9			dite	6

Figure 5a : Formes lexicales les plus spécifiques du voisinage d' "économie réelle" (seuil de spécificité de 6)

Candidat-sème	Spf	Candidat-sème	Spf	Candidat-sème	Spf
budget#subst	21	appréciable#adj	10	économique#adj	9
ressource#subst	16	capitaliste#adj	10	effondrement#subst	9
régir#v	15	collision#subst	10	enthousiasme#subst	9
argent#subst	14	contagion#subst	10	financier#adj	9
particulier#adv	14	décisif#adj	10	galaxie#subst	9
répercussion#subst	14	dysfonctionnement#subst	10	intense#adj	9
théâtre#subst	13	économie#subst	10	noeud#subst	9
bien#subst	12	époux#subst	10	pathologique#adj	9
chômage#subst	12	profond#subst	10	phénomène#subst	9
déterminant#adj	11	subit#adj	10	progressif#adj	9
diminution#subst	11	boursier#adj	9	retentissement#subst	9
néfaste#adj	11	craindre#v	9	rupture#subst	9
ralentissement#subst	11	D=dramaturgie	9	sous-production#subst	9
roi#subst	11	développement#subst	9	surproduction#subst	9

Figure 5b : Candidats-sèmes les plus spécifiques du voisinage d' "économie réelle" (seuil de spécificité de 9)

5.2 Filtrage par catégorie et émergence d'une forme sémantique

Les principales catégories choisies manuellement correspondent aux idées suivantes : la maladie ; le cataclysme ; le choc ou la brutalité ; la réalité ou, par opposition, la virtualité ; l'économie dans sa dimension matérielle. Les classes définies ont un degré de généralité variable. De plus, elles ne forment pas une partition : elles se superposent parfois et ne couvrent pas toutes les facettes sémantiques présentes dans l' "*économie réelle*". Certains candidats-sèmes sont donc affectés à plusieurs classes, tandis que d'autres ne rejoignent pas de classe particulière.

Pour constituer chaque catégorie, les listes de formes et de candidats-sèmes de spécificité supérieure à 2 ont été parcourues, avec un souci d'exhaustivité. L'affectation d'unités à certaines catégories s'est heurtée à des problèmes d'ambiguïté et à des cas d'incertitude. Des vérifications en contexte pour les formes et des recherches des formes génératrices pour les candidats-sèmes ont quelquefois été effectuées pour trancher sur l'affectation à une catégorie, mais cette procédure de contrôle n'a pu être systématisée, d'une part à cause d'usages variés des formes selon les contextes ou d'un trop grand nombre de formes génératrices, d'autre part en raison de la trop grande quantité de vérifications à faire.

A titre d'exemple, considérons les catégories suivantes : la catégorie 'maladie' (figure 6a) et la catégorie 'choc, brutalité' (figure 6b).

Trait	Spécif	Forme	Spécif
néfaste#adj	11	crise	9
contagion#subst	10	contagion	6
dysfonctionnement#subst	10	affectée	4
pathologique#adj	9	injectés	3
dépression#subst	8	affecter	3
trouble#subst	8	aggravée	3
crise#subst	7		
mal#subst	7		
physiologique#adj	6		
épidémie#subst	5		
infection#subst	5		
maladie#subst	5		
contagieux#adj	4		
remédier#v	4		
saignée#subst	4		
affecter#v	3		
bistouri#subst	3		
défaillir#v	3		
nuisible#adj	3		
psychose#subst	3		
soigner#v	3		

*Figure 6a : Catégorie **maladie** des unités spécifiques d' "économie réelle"*

Trait	Spécif	Forme	Spécif
effondrement#subst	9	choc	4
tarissement#subst	8	dommages	3
décrue#subst	8	éclate	3
dépression#subst	8	onde	3
choc#subst	7	fumée	3
violemment#adv	5	tempête	3
désagréger#v	4		
débris#subst	4		
déferler#v	3		
secousse#subst	3		
cataclysme#subst	3		
tempête#subst	3		
inondation#subst	3		

*Figure 6b : Catégorie **cataclysme** des unités spécifiques d' "économie réelle"*

Dans les deux cas, le nombre d'unités affectées à la catégorie est plus important sur le plan sémique que sur le plan lexical. De plus, certaines idées sensibles à la lecture mais sous-jacentes au niveau des formes lexicales apparaissent explicitement au niveau des candidats-sèmes. Par exemple, la maladie prend un caractère beaucoup plus prégnant et tangible avec

des candidats-sèmes tels que /pathologique/, /trouble/, /infection/, /épidémie/ ou encore /maladie/ ; de même, l'ébranlement et la violence liés à la crise, que seul *impact* reflète assez explicitement sur le plan lexical s'imposent avec force sur le plan sémique, avec des candidats-sèmes tels que /effondrement/, /heurt/, /brusque/, /violemment/ ou encore /secousse/. De façon générale, les catégories sont plus riches sur le plan sémique que sur le plan lexical, parce qu'elles contiennent plus de candidats-sèmes que de formes mais aussi, et surtout, parce que des idées perçues à la lecture sont exprimées clairement par les représentants sémiques alors qu'elles sont seulement sous-jacentes à travers les représentants lexicaux.

6 CONCLUSION

Cette étude décrit une procédure d'annotation en traits sémantiques de corpus, évaluée à travers la confrontation d'un corpus non annoté à son image annotée en candidats-sèmes. Les expériences réalisées indiquent une convergence entre l'évaluation intuitive de lecteur, le plan lexical et le plan sémique. Cette convergence se manifeste aussi bien à échelle globale (spécificités totales d'un journal par rapport à l'autre) que locale (focalisation sur le voisinage d'un mot-pôle). Les résultats valident ainsi la procédure d'annotation sémantique utilisée. Par ailleurs, l'approche en candidats-sèmes permet de faire émerger des formes sémantiques au voisinage d'un mot-pôle de façon plus marquée qu'au niveau lexical, d'une part en raison d'un accroissement des candidats-sèmes constitutifs de la forme sémantique, d'autre part en la profilant de façon plus fouillée que ne le fait le palier lexical de la forme présente.

L'enrichissement que propose l'annotation sémique est prometteur mais nécessite de se pencher sur le filtrage du bruit généré par l'annotation et sur le problème d'une polysémie inhérente à certains candidats-sèmes introduits. L'intégration d'informations domaniales ou encore la mise en place de représentations structurées des candidats-sèmes constituent des pistes susceptibles de réduire le problème. A travers ces développements, des perspectives plus larges s'ouvrent, comme la modélisation du sens pour la veille lexicale ou encore la détection de la néosémie.

7 RÉFÉRENCES

Brunet E. (2000). « Qui lemmatise dilemme attise ». *Scolia, 11e rencontres linguistiques en pays rhénan*, n°13, p. 7-32.

Dendien J. et Pierrel J.-M. (2003). « Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence ». *TAL*, 44/2, p. 11-37.

Grzesitchak M., Jacquey E. et Valette M. (2007). « Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies ». *ARCo'07 – Cognition, Complexité, Collectif.* Acta-Cognitica, p. 227-235.

Habert B. (2005). « Portrait de linguiste(s) à l'instrument ». *Revue Texto ! Textes et cultures*, vol. X, n°4, disponible sur http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html.

Loiseau S. (2006). *Sémantique du discours philosophique : du corpus aux normes. Autour de G. Deleuze et des années 60*. Thèse de Doctorat, Paris X-Nanterre.

Mayaffre D. (2008). « De l'occurrence à l'isotopie. Les cooccurrences en lexicométrie ». *Textes, documents numériques, corpus. Pour une science des textes instrumentée, Syntaxe & Sémantique*, 9, p. 53-74.

Mellet S. (2002). « Lemmatisation et encodage grammatical : un luxe inutile ? ». Lexicometrica, 3, 12.

Rastier F. (2001). *Arts et sciences du texte.* Paris : PUF.

Salem A., Lamalle C., Martinez W., Fleury S., Fracchiolla B., Kuncova A. et Maisondieu A. (2003). *Lexico3 – Outils de statistique textuelle. Manuel d'utilisation.* Syled-CLA2T, Université de la Sorbonne nouvelle – Paris 3 http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW.

Slodzian M. (1999). « WordNet et EuroWordNet – Questions impertinentes sur leur pertinence linguistique ». *Sémiotiques,* n°17, p. 51-70.

Valette M., Estacio-Moreno A., Petitjean E. et Jacquey E. (2006). « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens ». *Verbum ex machina* (TALN 06), P. Mertens, C. Fairon, A. Dister, P. Watrin (éds). Cahiers du CENTAL, 2.1, UCL Presses Universitaires de Louvain. Volume 1. p. 357-366.

Valette M. (2008). « A quoi servent les lexiques sémantiques ? Discussion et proposition ». Dans M. Constant et al. (®ds.), Description linguistique pour le traitement automatique du français, *Cahiers du CENTAL*, n°5 – décembre 2008, PUL, p. 43-58.

CONSTITUTION D'UN CORPUS D'ERREURS DU DACTYLOGRAPHE

Agnès Souque
LIDILEM – Université Stendhal – Grenoble

RÉSUMÉ

La rédaction de textes sur ordinateur est aujourd'hui difficilement dissociable de l'utilisation d'outils de correction automatique. Si certains, comme les correcteurs orthographiques, sont relativement efficaces, d'autres comme les correcteurs grammaticaux sont encore limités par leur principe de fonctionnement. Pour résoudre ce problème, nous avons décidé de concevoir un outil de correction grammaticale fondé sur des principes différents. Dans cette optique, nous avons choisi une approche corpus pour nous permettre de réaliser des analyses linguistiques des erreurs des dactylographes et ainsi mieux les détecter. Malgré des contraintes importantes, nous avons constitué un corpus à partir de textes contenant des erreurs, que nous avons annotées en nous appuyant sur une typologie d'erreurs *ad hoc*. Les premières analyses, qui nécessitent d'être complétées et affinées, permettent déjà de mettre en évidence les principales erreurs commises dans les textes tapuscrits.

1 INTRODUCTION

Lorsque nous rédigeons, nous sommes amenés à faire des écarts par rapport aux normes de grammaire et d'orthographe. Des outils informatiques existent aujourd'hui pour nous aider à corriger ces écarts lorsque nous rédigeons sur un ordinateur. Ils sont capables, de manière plus ou moins fiable, de détecter aussi bien des erreurs d'orthographe que de grammaire. La correction grammaticale automatique du français, dans la catégorie des outils libres, est encore relativement limitée. Après avoir défini ce qu'est la correction grammaticale, nous verrons les faiblesses des outils actuels et proposerons des solutions pour rendre efficace la correction dans la très grande variété des documents dactylographiés. Nos solutions impliquent le développement d'un nouvel outil de correction. Dans ce but, nous avons besoin d'étudier les erreurs que commettent les dactylographes. Pour y parvenir, nous avons choisi une approche corpus, mais nous avons été confrontée à des contraintes qui remettent en cause la représentativité de notre corpus. Nous avons tout de même pu recueillir de nombreuses données dont l'annotation des erreurs nous a permis d'une part de définir une typologie des erreurs des dactylographes, d'autre part d'avoir un premier aperçu des principales erreurs qu'ils commettent.

2 RÉSOUDRE LE PROBLÈME DE LA CORRECTION GRAMMATICALE AUTOMATIQUE

L'informatique est aujourd'hui capable de nous aider dans la rédaction de nos textes. Cependant nous ne faisons pas toujours la distinction entre erreur de grammaire et erreur d'orthographe, et il est souvent reproché aux correcteurs orthographiques de ne pas détecter par exemple les erreurs d'accords. Nous commencerons donc par définir les notions de

grammaire et d'orthographe, telles que le linguiste les considèrent, mais également telles qu'elles sont traitées par les outils informatique.

Lorsque nous utilisons des outils capables de gérer les erreurs de grammaire, nous leur reprochons alors soit leur manque d'efficacité, soit leur excès de détections erronées. Nous verrons donc dans un second temps comment fonctionnent ces outils, quelles sont leurs faiblesses et quelles solutions nous pouvons apporter.

Pour finir, nous tenterons de donner un aperçu du vaste champs des documents que la correction grammaticale est susceptible de concerner.

2.1 Qu'est-ce que la correction grammaticale automatique ?

2.1.1 LA GRAMMAIRE EN LINGUISTIQUE ET EN INFORMATIQUE

En linguistique, la grammaire désigne "*l'étude systématique des éléments constitutifs et du fonctionnement [...] d'une langue"*. (Grévisse, 1993). La grammaire descriptive étudie la structure des phrases ainsi que les relations que les mots entretiennent entre eux, alors que la grammaire normative, celle qui est enseignée aux apprenants, fixe les règles de bonne formation des énoncés. Ainsi, un verbe qui n'est pas accordé correctement avec son sujet, l'utilisation d'un mauvais pronom relatif, le non respect de l'ordre correct des mots par exemple sont considérés comme des erreurs de grammaire. Cependant, pour certaines de ces erreurs, on parle également d'orthographe, et plus précisément d'orthographe grammaticale, orthographe de règle ou orthographe d'accord. Elle correspond aux modifications des mots en fonction de leur rôle dans la phrase. Elle régit principalement les accords entre les mots et détermine donc, entre autres, les marques du pluriel ou du féminin, les désinences de conjugaison, et implique de savoir identifier les rapports qu'entretiennent les mots d'une phrase entre eux afin de pouvoir les accorder.

L'orthographe grammaticale est à distinguer de l'orthographe lexicale, ou orthographe d'usage, qui correspond à la manière d'écrire les mots tels qu'indiqués dans les dictionnaires, sans que soit prise en compte leur fonction dans la phrase. Chaque mot a une orthographe définie, qui ne dépend pas de la grammaire, mais uniquement du lexique. Pour connaître cette graphie, il faut avoir recours à un dictionnaire (Bonnard, 1981). Une faute d'orthographe lexicale correspond à une différence de graphie d'un mot, par rapport à celle définie dans les dictionnaires. Les mots dits invariables ne sont concernés que par l'orthographe lexicale, de même que le radical des mots variables. En revanche, les variations dont sont sujets certains mots, comme les marques flexionnelles, sont du ressort de l'orthographe grammaticale.

D'un point de vue linguistique, "**toujoures*" est donc du domaine du lexique, mais "**j'ai écris*" et "**ils dorments*" sont des erreurs grammaticales.

En informatique, les notions de grammaire et d'orthographe ne sont pas exactement les mêmes qu'en linguistique. La distinction entre les deux types d'erreurs se fait en fonction du correcteur automatique qui est capable de les détecter. Le mot "orthographe" a un sens plus restreint qu'en linguistique. Il se réfère toujours à la graphie des mots, mais sans qu'il soit question de distinction entre les erreurs sur la graphie des radicaux par exemple, qui sont d'ordre lexical, et les erreurs sur la graphie des désinences qui sont de l'ordre du grammatical, d'un point de vue linguistique. Ainsi, tout mot dont la graphie n'existe pas dans la langue est considéré comme mal orthographié. Les erreurs grammaticales regroupent alors tous les autres écarts d'écriture, qui conduisent à des mots qui existent bels et bien mais qui ne sont pas en adéquation avec leur contexte syntaxique (mauvaise flexion, confusion de radical, etc.)

Ainsi, d'un point de vue informatique, les exemples "**toujoures*" ou "**ils dorments*" sont des erreurs d'orthographe, mais "**j'ai écris*" est une erreur grammaticale.

Linguistes et informaticiens n'ont donc pas tout à fait la même notion de l'orthographe et de la grammaire. Du point de vue de l'un ou de l'autre, certaines erreurs ne sont pas classées dans la même catégorie. Il y a un chevauchement des niveaux orthographique et grammatical, en informatique et en linguistique, que nous représentons dans la Figure 1par la zone hachurée.

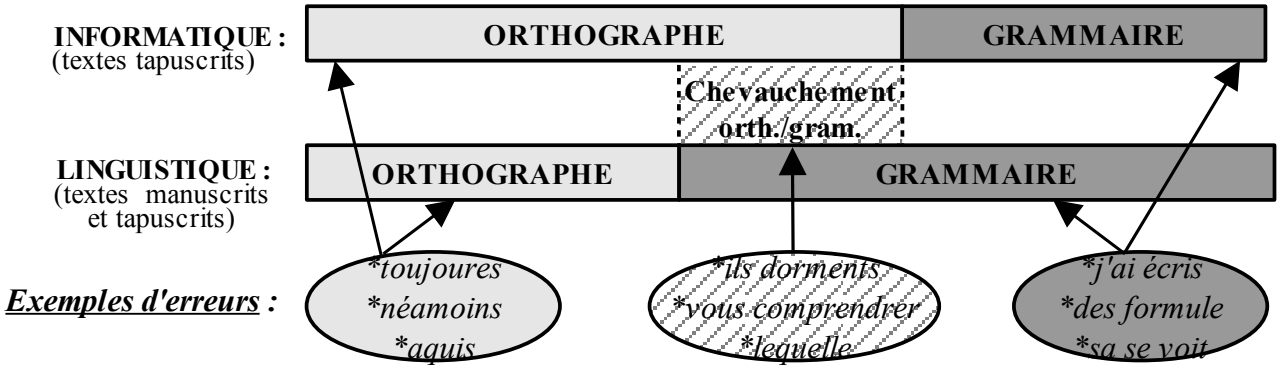


Figure 1 : Chevauchement de l'orthographe et de la grammaire, en linguistique et en informatique

2.1.2 CORRECTION GRAMMATICALE VS. CORRECTION ORTHOGRAPHIQUE

La grammaire et l'orthographe peuvent être vérifiées de manière automatique lors de la rédaction de documents sur un ordinateur. Pour cela il existe des correcteurs orthographiques et grammaticaux. Les correcteurs orthographiques sont les plus répandus. Ils fonctionnent à partir d'un lexique de formes fléchies, c'est-à-dire de toutes les flexions des mots, et non à partir des seules entrées des dictionnaires, qui sont constituées par les seuls lemmes. Lorsqu'ils rencontrent un mot du texte qui ne figure pas dans leur lexique, ils signalent une erreur d'orthographe. Ils ne tiennent pas compte du contexte dans lequel ce mot se trouve. Il serait d'ailleurs plus approprié de parler d'erreur de graphie, ou d'épellation, au lieu d'orthographe. L'anglais utilise d'ailleurs le terme *spellchecker* (littéralement "vérificateur d'épellation") pour désigner ce que nous appelons maladroitement "correcteur orthographique".

Les correcteurs grammaticaux ont pour vocation de corriger, ou plutôt de détecter et signaler les erreurs de grammaire, c'est-à-dire toutes les erreurs qui ne sont pas repérées par les correcteurs orthographiques. Comme nous le verrons par la suite, ils ne sont en réalité capables de traiter qu'une partie de ces erreurs.

2.1.3 DÉFINITIONS DE CORRECTION ET ERREUR GRAMMATICALES

Nous utilisons les termes "correcteur" ou "correction" tout au long de cet article pour parler des outils automatiques car ce sont les termes qui sont communément utilisés en informatique pour désigner ces outils. Ils ne sont cependant pas tout à fait appropriés. En effet, les outils recherchent les erreurs et font éventuellement des suggestions de correction, que l'utilisateur est libre d'accepter ou pas, mais en aucun cas ils ne corrigent automatiquement les erreurs. Ce que nous nommons correcteur ou correction automatique désigne donc en fait la vérification ou la détection automatique des erreurs.

Par ailleurs, nous avons montré plus haut qu'il existe un chevauchement entre la grammaire et l'orthographe selon le point de vue informatique ou linguistique. Lors de l'annotation des erreurs de notre corpus, nous y avons été confrontée. Notre formation de linguiste nous incite à étiqueter les erreurs en suivant la distinction grammaire/orthographe linguistique, de manière à pouvoir par exemple comparer nos analyses par la suite avec les autres analyses d'erreurs qui ont pu être réalisées en linguistique. Cependant, nous avons constitué notre

corpus dans le cadre de la correction automatique, afin notamment d'établir les règles adéquates pour détecter les erreurs que nous souhaitons corriger. Dans cette optique, il nous a semblé évident que les erreurs de notre corpus devaient être étiquetées en nous appuyant sur le point de vue informatique de la grammaire et l'orthographe.

2.2 Quels sont les problèmes de la correction grammaticale automatique ?

2.2.1 FONCTIONNEMENT DES OUTILS

Parmi les outils de correction grammaticale existants, nous pouvons distinguer deux catégories : les logiciels propriétaires d'une part, tels Cordial ou Antidote, et les logiciels libres d'autre part, tels LanguageTool ou An Gramadoír. Le fonctionnement des logiciels propriétaires n'étant pas documenté, pour des raisons évidentes de secret industriel, nous n'avons pas de moyen d'étudier la structure de ces outils. En revanche, nous savons comment fonctionnent les correcteurs libres sur lesquels nous avons travaillé. Ils ont une structure similaire, dans laquelle les traitements des données s'enchainent de manière séquentielle (Figure 2)

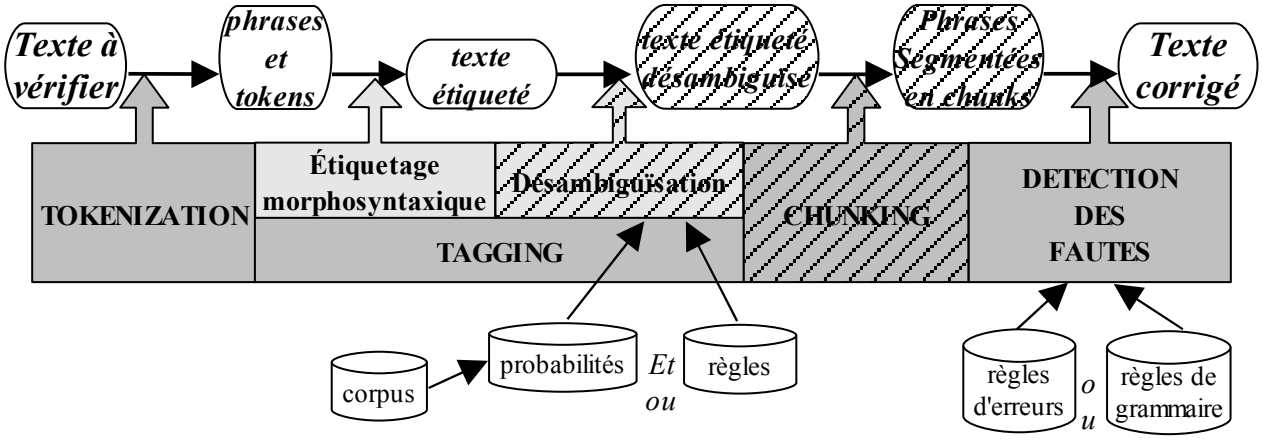


Figure 2 : Structure générale d'un correcteur grammatical
(les zones hachurées correspondent aux traitements facultatifs)

Le texte est d'abord segmenté en phrases et en *tokens* (mots et ponctuations). C'est la tokenization.

Chaque *token* est ensuite étiqueté morphologiquement et reçoit autant d'étiquettes que de catégories ou sous-catégories grammaticales auxquelles il appartient. Environ 46% des mots du français sont en effet ambigus. Un mot tel que *"passante"* aura ainsi trois étiquettes :

Nom - féminin singulier

Adjectif – féminin singulier

Verbe – participe présent

Suite à l'étiquetage, certains correcteurs effectuent une désambiguïsation afin de limiter le nombre d'étiquettes des mots ambigus. Certains correcteurs utilisent une méthode probabiliste, d'autres les règles manuelles, certains combinent les deux méthodes, enfin d'autres ne font pas du tout de désambiguïsation.

L'étape suivante concerne peu de correcteurs mais nous la mentionnons tout de même car elle nous intéresse particulièrement pour l'outil que nous souhaitons développer. Il s'agit du découpage en syntagmes minimaux, ou *chunking*, qui permet de travailler à un niveau intermédiaire entre la phrase et le mot, en définissant des groupes de mots. Ce type de

segmentation peut être très utile dans le traitement automatique des langues (Abney, 1991) comme nous le verrons par la suite.

La dernière étape est celle de la détection des erreurs de grammaire. Il existe là encore deux manières de procéder, toutes deux fondées sur le principe du *pattern-matching*, c'est-à-dire sur la correspondance exacte entre un segment du texte et un modèle. Certains correcteurs utilisent des règles de grammaire. Elles décrivent des modèles (*patterns*) grammaticalement corrects. Si une partie du texte ne correspond à aucun *pattern*, une erreur est détectée. Les correcteurs que nous avons analysés utilisent au contraire des règles d'erreurs. Ils comparent alors le texte non pas à des modèles corrects, mais à des modèles de fautes. Une erreur est détectée lorsque le texte et le *pattern* coïncident.

2.2.2 PROBLÈMES IDENTIFIÉS

Les travaux sur An Gramadoír (Lechelt, 2005) ainsi que ceux que nous avons effectués sur LanguageTool (Souque, 2008) nous ont permis d'identifier les limites de ce types de correcteurs. La principale limite est due à l'utilisation du principe du *pattern-matching*, dont les conséquences sont multiples.

Pour être détectée, une erreur doit être décrite dans une règle. Il faut ainsi prévoir toutes les erreurs possibles pour les décrire dans des règles et pouvoir les détecter. Pour les erreurs d'accord au sein d'un syntagme nominal par exemple, il faut prévoir tous les cas en fonction des combinaisons possibles des mots et des genres et nombres auxquels ils peuvent être déclinés. On obtient alors une explosion combinatoire du nombre de règles à créer, simplement pour gérer les accords nominaux. Il est évident de manière plus générale qu'il n'est pas possible d'écrire une règle pour chaque erreur potentielle et que certaines erreurs ne sont donc pas détectées. C'est également le cas lorsque des mots sont mal orthographiés, mal étiquetés, oubliés, ajoutés ou mal ordonnés. Ils gênent alors la bonne application des règles d'erreurs en empêchant la correspondance entre la séquence du texte où il se trouve et le modèle d'erreur. Le *pattern-matching* est donc générateur de silence dans la détection de fautes, mais également de bruit. En effet, le grand nombre de règles nécessaires qu'il impose conduit à des détections en cascade à partir d'une seule erreur, soit parce que plusieurs règles s'appliquent à cette seule erreur, soit parce que l'erreur en question provoque l'application de règles à tort sur les mots suivants. Le *pattern-matching* implique aussi la limitation des détections d'erreur au contexte immédiat. Il n'est pas possible de prévoir des règles tenant compte de tout ce qui peut se trouver entre deux éléments distants devant s'accorder. Enfin, la correction grammaticale ainsi formalisée est victime d'un cercle vicieux. Pour que les règles s'appliquent correctement, elles doivent reconnaître des séquences de textes telles qu'elles les décrivent. Cependant, lorsque des éléments du texte sont erronés, ils peuvent entraîner la non reconnaissance de ces séquences par les règles. Ainsi, pour que la détection d'erreurs soit la plus efficace, le texte ne doit pas contenir d'erreurs.

2.2.3 SOLUTIONS PROPOSÉES

Il est indispensable, afin d'améliorer la correction grammaticale, de simplifier les règles et de contourner le *pattern-matching*. Dans ce but, nous proposons une approche "gauche-droite" dont le principe est de construire simultanément l'analyse morphosyntaxique et la correction au fur et à mesure de la rédaction ou la lecture de la phrase, et de rechercher des incohérences grammaticales au lieu d'énumérer toutes les erreurs possibles.

La recherche d'incohérences implique d'une part la segmentation en *chunks* que nous évoquions en 2.1.1. et le principe d'unification de structures de traits (Abeillé, 1993), d'autre part le principe des latences (Tesnière, 1959) ou d'attentes réciproques (Lebarbé, 2002)

Un *chunk* est un groupe de mots au sein duquel les mots entretiennent des relations de dépendance et dont la structure interne est relativement figée. Les contraintes d'accord entre les mots sont alors assez fortes, ce qui nous intéresse particulièrement pour la correction grammaticale.

Les *chunks* sont délimités grâce aux mots grammaticaux, à la ponctuation ou aux marques morphologiques, et sont donc relativement faciles à définir. En général, un *chunk* commence par un mot grammatical ou juste après une ponctuation, et se termine juste avant une ponctuation ou le mot grammatical suivant, qui marquent alors le début d'un nouveau *chunk*. Ainsi, la définition d'un *chunk* ne se fait pas en fonction de son contenu, mais en fonction de ses marqueurs de début et de fin.

Exemple de segmentation : *[*Les personnes] [en situation] [de test] [s'autosurveille énormément]*

Il existe également des relations de dépendance *inter-chunks*. Ces propriétés de dépendance vont s'avérer très utiles pour vérifier les accords entre différents syntagmes.

Combinée à la segmentation en *chunks*, l'unification de structures de traits va permettre la détection de toutes les erreurs d'accord en évitant l'utilisation des milliers de règles nécessaires avec le *pattern-matching.*

« L'unification de deux structures de traits A et B (notée A ∪ B) est la structure minimale qui est à la fois une extension de A et de B. Si une telle extension n'existe pas, l'unification «échoue» (ce qui est noté ⊥) ».(Abeillé, 1993)

Les structures de traits décrivent chaque élément d'une phrase en énumérant ses caractéristiques linguistiques, syntaxiques ou sémantiques, sous la forme de liste de couples trait-valeur. Les étiquettes des mots en sont un exemple. Les informations contenues dans les étiquettes correspondent aux traits. L'unification consiste à combiner les traits de deux étiquettes et à vérifier leur compatibilité. Deux structures de traits peuvent s'unifier si elles ont les mêmes valeurs pour des traits identiques. Par exemple, l'unification n'est pas possible entre l'étiquette du nom "*genoux"* et celui du déterminant "*le"*, car la valeur du trait nombre n'est pas identique (Figure 3).

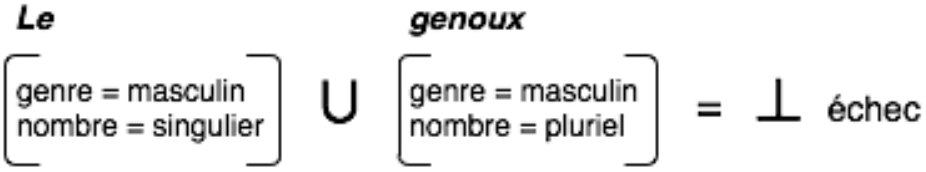


Figure 3 : Exemple d'échec d'unification

En utilisant l'unification de structure de traits, la détection des erreurs d'accord au sein d'un *chunk* est possible, même si un mot est ambigu ou mal étiqueté. Les *chunks* délimitent donc des zones de calcul au sein desquelles l'unification permet de vérifier que tous les éléments s'accordent bien entre eux. De plus, en attribuant des traits au *chunks*, ils peuvent alors s'unifier entre eux, et des erreurs entre groupes de mots peuvent ainsi être détectées, comme par exemple l'accord entre le groupe nominal sujet et le groupe verbal.

En complément dans la détection des incohérences grammaticales, le principe des latences (Tesnière, 1959) permet, à partir de règles d'attentes (un déterminant attend un nom par exemple), de signaler les éléments inattendus. La vérification grammaticale est donc effectuée dans l'ordre de lecture des *tokens* par la validation des attentes réciproques, la délimitation des

chunks et les tests d'unification : l'échec de l'un des trois constituant une détection d'incohérence grammaticale et donnant son explication. Cette approche implique une reconsidération complète des formalismes existants en ne déclarant non plus des erreurs, mais ce qui est attendu.

2.3 À quels documents s'applique la correction grammaticale ?

2.3.1 DOCUMENTS DACTYLOGRAPHIÉS

Les outils de vérification automatique de la grammaire sont destinés à analyser tout texte dactylographié numérique. Ce type de texte peut être généré de manières diverses. La plus commune est sans aucun doute la saisie au clavier dans les logiciels divers permettant l'édition, tels les traitements de textes, les clients email ou les navigateurs internet.

Une autre manière de générer du texte dactylographié implique l'utilisation de logiciels de reconnaissance optique de caractères (ou OCR). Ces logiciels permettent de numériser des textes sur support papier pour permettre leur édition dans un logiciel de traitement de texte. Selon la qualité et le type du texte numérisé, la fiabilité de la reconnaissance de caractères peut être très aléatoire. En plus des éventuelles erreurs d'écriture contenues dans le texte original, le traitement par un logiciel d'OCR peut ajouter de nombreuses nouvelles erreurs dues à des défauts de reconnaissance de caractères. Il est évident que la correction automatique est indispensable avec ce type de logiciels dans lesquels elle est d'ailleurs généralement intégrée. Cependant, nous avons préféré laisser de côté les documents issus de logiciels d'OCR qui constituent un type à part, pour nous concentrer sur les textes rédigés directement à l'aide d'un clavier dans tout logiciel permettant la rédaction. Il existe également des logiciels de reconnaissance automatique de la parole (ou ASR) qui permettent de retranscrire à l'écrit les paroles prononcées par l'utilisateur. Ce type d'outil est également générateur de nombreuses erreurs, principalement dues à des défauts de reconnaissance des mots prononcés. De même que pour les logiciels d'OCR, nous laisserons ce type de logiciels de côté. Nous nous concentrons donc sur les textes qui sont saisis directement à partir d'un clavier d'ordinateur.

2.3.2 DOCUMENTS VARIÉS

Les types de textes que nous sommes amenés à taper sur un ordinateur et dont nous pouvons avoir besoin de vérifier la correction sont de nature très variée. Nous en dressons une liste non exhaustive qui comprend :

- des documents de correspondance : ils peuvent être rédigés avec des niveaux de formalité variés, par des personnes de tranche d'âge et de niveau d'études différents, mais sans contrainte de temps. Ce type de documents est relativement difficile à obtenir pour notre corpus. D'une part ils sont souvent manuscrits, ce qui ne nous intéresse pas, d'autre part ils sont relativement personnels et privés et donc rarement librement disponibles.
- des documents rédigés dans un cadre professionnel : qu'il s'agisse de rapports, de comptes rendus, de notes, etc.., le niveau de formalité est généralement assez élevé et la contrainte temporelle plutôt faible. Les erreurs présentes dans de tels documents dépendent beaucoup du niveau d'étude et du poste occupé. Mais ces documents sont souvent régis par le secret professionnel, ce qui les rend difficilement accessibles.
- des documents rédigés dans un cadre scolaire : ils présentent presque toujours un niveau de formalité important puisqu'ils sont généralement associés à une évaluation. Ce qui influera le plus sur la quantité et les types de fautes sera bien sûr l'âge et le niveau d'étude, mais aussi le niveau de contrainte temporelle.

- des textes issus d'internet : il est souvent difficile, voire impossible, de recueillir des informations sur les scripteurs de documents en ligne. Les erreurs dans les textes dépendent surtout des auteurs et du sujet du site, ce dernier impliquant souvent un degré de formalité plus ou moins important.

2.3.3 SCRIPTEURS VARIÉS

Les personnes qui saisissent des textes sur ordinateur et sont donc amenées à utiliser la correction grammaticale automatique sont de profils très variés. Elles sont de toutes tranches d'âges et de tous niveaux d'études. Le niveau d'étude est un paramètre important car il est généralement corrélé à la fréquence des erreurs. Plus il est élevé, plus les connaissances en grammaire sont sollicitées, et moins les erreurs sont nombreuses. En effet, (Lucci et Millet, 1994) apportent la preuve chiffrée que "*la fréquence et le type des variantes orthographiques produites par un scripteur dépendent crucialement de son niveau d'étude et de sa profession*". Pour refléter l'usage général de la langue, notre corpus doit donc comporter des écrits de tous niveaux aussi bien de personnes n'ayant pas le Bac que de personnes ayant suivi des études supérieures.

Les scripteurs peuvent également être des francophones natifs ou bien des apprenants de la langue à des niveaux d'avancement divers. Pour ce qui est des apprenants de français langue étrangère ou seconde, ils constituent un public important potentiellement utilisateur de correcteurs grammaticaux, et il nous parait donc indispensable d'inclure dans notre corpus des écrits émanant de ce type de scripteurs. Cependant, les erreurs commises par les apprenants sont différentes de celles commises en langue maternelle. Certaines de ces erreurs sont difficilement détectables par les outils de correction actuels et nécessitent des programmes spécifiquement adaptés (Granger, 2007), à but pédagogique notamment avec des rétroactions adaptées, ce qui n'est pas le cas de l'outil que nous souhaitons développer. Nous n'excluons donc pas les textes rédigés par des auteurs non natifs, mais ils seront en proportion moindre dans notre corpus, par rapport aux textes de francophones natifs.

3 LE CHOIX D'UNE APPROCHE CORPUS

Notre besoin d'analyser les erreurs des dactylographes nous a naturellement conduite vers une approche corpus, approche communément choisie dans le domaine du traitement automatique des langues où la nécessité de disposer de données réelles pour réaliser des analyses linguistiques est incontestable. Nous avons cependant vite dû faire face à de nombreuses contraintes dans la constitution du corpus, contraintes mettant à mal la possibilité d'obtenir un corpus représentatif des erreurs dans les documents dactylographiés.

3.1 Pourquoi une approche corpus ?

3.1.1 BESOIN DE DONNÉES RÉELLES

Pour être efficace, un correcteur grammatical automatique doit être capable de détecter un maximum des erreurs d'orthographe et de grammaire que commettent les scripteurs. Il doit donc être conçu de manière à savoir reconnaître ces écarts d'écriture. Il est ainsi nécessaire lors de sa conception de disposer d'une typologie des erreurs possibles afin qu'il ait les compétences requises pour les détecter. Pour élaborer cette typologie, une analyse linguistique des erreurs est nécessaire.

« *Le choix d'un champ d'étude composé essentiellement de textes proprement littéraires [...] correspond aussi au désir, ressenti par de nombreux linguistes, d'échapper aux pièges d'une syntagmatique purement locale, génératrice de contre-exemples artificiels.*

« *Une telle approche nous semble aussi naturelle que celle du zoologiste préférant l'étude des espèces 'réelles' à celle des animaux fabuleux ou mythologiques.* » [Formalismes pour l'analyse et la synthèse de textes littéraires, Paul Braffort, dans Atlas de Littérature Potentielle (OULIPO)]

L'outil étant destiné à analyser des textes authentiques, contenant des erreurs non artificielles, son moteur de correction doit donc s'appuyer sur des données réelles. La nécessité de travailler sur des données attestées pour une analyse linguistique mène généralement à l'utilisation de corpus. Nous avons donc choisi cette approche pour analyser les erreurs d'écriture et en élaborer une typologie pour la conception de notre correcteur.

3.1.2 ABSENCE DE CORPUS ADAPTÉ

Il n'existe pas à l'heure actuelle de large corpus d'erreurs disponible. Le corpus FRIDA, constitué dans le cadre d'un projet de développement d'un logiciel d'ELAO, est relativement important avec plus de 450 000 mots. Cependant, pour plusieurs raisons, ce corpus ne correspond pas à ce dont nous avons besoin. Tout d'abord il est constitué uniquement de productions d'apprenants du français langue étrangère. Même si cette catégorie de locuteurs du français n'est pas à ignorer dans notre corpus, celui-ci ne serait en aucun cas représentatif de la variété des utilisateurs de logiciels de rédaction et d'outils de correction s'il ne contenait que des écrits de francophones non natifs.

Mais surtout, les textes constituant le corpus FRIDA sont initialement rédigées sur papier, puis numérisés, ce qui pose problème pour nous. Rien ne tend en effet à prouver que les erreurs commises dans les textes manuscrits sont les mêmes que dans les textes dactylographiés. Au contraire, la configuration spatiale des touches des claviers peut "*être à l'origine d'erreurs que l'on n'observerait pas en production manuscrite.[...]En écriture sur clavier, à ces erreurs* (manuscrites) *(toujours possibles) s'ajoutent celles qui peuvent émaner de la proximité spatiale de certaines lettres, et ce, même si « lettres substituantes » et « lettres substituées » n'ont rien en commun dans le système alphabétique. Par exemple, les touches correspondant aux lettres E, R, S, D, F sont toutes situées sur la même zone d'un clavier AZERTY. Mais elles n'ont aucun point commun en termes de graphie (comme par exemple p et q), ou de sonorité (comme m et n)."* (Boissière, 2007).

Par ailleurs, les textes originaux manuscrits ont été retranscrits sur clavier. Nous pouvons donc émettre l'hypothèse que des erreurs supplémentaires ont été rajoutées par inadvertance, avec par exemple des fautes de frappe dues à la configuration spatiale du clavier comme évoqué plus haut, ou bien que des erreurs authentiques ont été inconsciemment corrigées de manière automatique par les retranscripteurs. Ceci introduit donc un biais dans l'authenticité des erreurs d'écriture commises. Nous sommes cependant consciente du fait que les versions numériques des textes du corpus Frida sont tout de même des textes dactylographiés qui contiennent des erreurs et pourraient donc correspondre à nos besoins. Mais parce qu'il n'est pas possible de discerner les erreurs manuscrites des réelles erreurs dactylographiques, nous préférons les écarter.

Pour des raisons similaires, nous avons également écarté le corpus d'erreurs (COVAREC, 1994). Contrairement au corpus FRIDA, il est constitué de textes de francophones natifs, mais ce sont des documents manuscrits.

3.2 Quelles sont les contraintes ?

3.2.1 LES CONDITIONS DE SCRIPTIONS

Notre corpus doit refléter des degrés de formalité divers, qui influent sur l'attention portée par le scripteur à ses écrits. Le plus haut niveau de formalité concerne les situations où l'écrit a

pour destinataire une tierce personne, dans un rapport hiérarchisé, avec pour objectif d'obtenir quelque chose, comme un emploi avec une lettre de candidature, ou une bonne note à un devoir. Des copies d'étudiants (devoir, exposé, examen, mémoire, etc.) ou des documents rédigés dans un cadre professionnel (rapport, lettre, etc.) font donc partie des situations de rédaction avec un haut degré de formalité qui peuvent nous intéresser. Au contraire, si l'écrit est destiné à une personne proche, s'il n'y a pas de relation hiérarchisée, s'il n'y a pas d'enjeu particulier, le niveau de formalité est faible et l'attention portée à la qualité de la rédaction beaucoup plus relâchée. Les erreurs se trouvent alors généralement en plus grand nombre. Cela peut être le cas dans des mails, les discussions instantanées, certains blogs, etc. L'auto-surveillance est minimale lorsque le scripteur et le lecteur ne sont qu'une seule et même personne, c'est-à-dire lorsque l'écrit n'est destiné qu'à soi, comme dans le cas de prises de notes par exemple. Le niveau de formalité étant nul, le risque de trouver de nombreuses erreurs est beaucoup plus élevé.

Nous souhaitons également représenter dans notre corpus différents niveaux de contrainte lors de la rédaction. Ces niveaux de contrainte correspondent au temps dont dispose le scripteur pour rédiger son texte. Plus il est réduit, plus le risque d'avoir des erreurs est important. En situation d'examen par exemple, un temps limité pour la rédaction ne permet pas à l'étudiant d'effectuer toutes les relectures nécessaires pour vérifier son texte. Il laisse ainsi des erreurs de grammaire qu'il aurait sans doute corrigées en l'absence de contrainte temporelle. Dans d'autres situations, c'est le temps disponible pendant l'écriture qui est très limité. Lors de la prise de notes, le scripteur dispose de très peu de temps pour réfléchir à l'orthographe, à la grammaire. Il est alors amené à commettre beaucoup plus d'erreurs.
Il nous faudrait donc pour notre corpus des textes rédigés dans un contexte formel ou libre, ainsi qu'avec ou sans contrainte de temps.

3.2.2 L'IDENTIFICATION DES AUTEURS

Nous souhaitons identifier l'auteur de chacun des textes afin d'utiliser les informations accordées par l'auteur comme filtre de perception du corpus. Nous souhaitons notamment connaître son âge, son sexe, son niveau d'étude et sa langue maternelle, qui sont autant de paramètres pouvant influer sur la quantité et le type des erreurs commises.

Sur Internet, les scripteurs sont souvent difficiles, voire impossibles à identifier. Les sites officiels ou professionnels ont par exemple généralement plusieurs auteurs, ce qui complique d'autant la tâche d'identification du scripteur de textes potentiellement intéressants. Les auteurs des messages postés dans les forums de discussion ou en commentaires de certains sites sont malheureusement aussi presque impossibles à identifier pour obtenir des informations, du fait de l'usage de pseudonymes.

Les catégories de textes dont nous pouvons identifier les auteurs sont ainsi relativement réduites, ce qui complique notre travail de recueil du corpus.

3.2.3 L'OBTENTION DES DOCUMENTS

Pour des raisons juridiques évidentes, nous avons besoin de l'autorisation de l'auteur pour pouvoir utiliser son texte. Or comme nous venons de l'évoquer, les auteurs sont souvent difficiles à contacter, et lorsqu'ils peuvent l'être, peu répondent positivement à la demande d'autorisation, soit par désintérêt, soit par méfiance.

A l'inverse, pour beaucoup de documents nous intéressant l'auteur est très facilement identifiable et disposé à donner son accord. Il s'agit des documents ou correspondance de proches (amis et famille). Cependant, par souci éthique, nous avons décidé de ne pas les inclure à notre corpus.

Nous avions également à disposition de nombreux textes dactylographiés, mais sous format papier. Les auteurs pouvaient être identifiés, mais il aurait fallu numériser ces textes, soit par OCR, soit par saisie au clavier, pour pouvoir en faire l'analyse. Comme nous l'avons indiqué précédemment, la retranscription de textes peut conduire à l'ajout ou la correction involontaire d'erreurs, et altère alors l'authenticité de ces erreurs. Nous n'avons donc pas utilisé ces documents.

3.3 Quelle représentativité peut-on attendre ?

Un corpus est *"une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage"* (Sinclair, 1996, trad. Habert et al., 1997). Nous avons collecté des textes, dont la principale caractéristique est de contenir des erreurs, dans le but d'avoir un échantillon des écarts d'écriture tapuscrits. En ce sens, notre collection de textes peut-être assimilée à un corpus. Cependant, une des caractéristiques d'un corpus est d'être un échantillon représentatif de la population globale qu'il étudie. Cette représentativité est en fait difficilement atteignable pour notre corpus.

3.3.1 D'INNOMBRABLES VARIÉTÉS DE DOCUMENTS

Notre corpus est censé représenter les erreurs commises par les scripteurs sur ordinateur. Cependant, la très grande variété aussi bien des types de documents que des scripteurs nous amène à conclure qu'il n'est pas possible d'intégrer à notre corpus une représentation de tous ces écrits.

Finalement nous pouvons dire que nous avons collecté des textes représentatifs d'un échantillon de documents et de scripteurs. A supposer que nous collections des documents dont les auteurs ne seraient pas identifiés, afin de limiter les les difficultés à recueillir les textes et les contraintes auxquelles nous avons dû faire face, certains types de documents nous resteraient inaccessibles, comme par exemple les documents régis par le secret professionnel, ou encore les documents personnels non diffusés.

Notre échantillon n'est donc pas représentatif de l'ensemble des écrits dactylographiés, mais il va cependant nous permettre d'établir une première typologie des erreurs tapuscrites, que nous pourrons faire évoluer par la suite si d'autres types de documents viennent compléter notre corpus.

3.3.2 DES BIAIS DANS LES MÉTHODES D'ACQUISITION DES DOCUMENTS

Devant la difficulté à trouver des textes pouvant convenir pour notre corpus, nous nous sommes principalement focalisée sur les scripteurs et les documents du milieu dans lequel nous évoluons, à savoir le milieu universitaire, notamment en informatique et traitement automatique des langues. La quasi totalité de notre corpus est ainsi constitué de textes d'étudiants, limitant drastiquement la variété des scripteurs.

Par ailleurs, pour obtenir un nombre important de documents, nous avons mis en place un protocole de saisie de dictées, toujours par des étudiants. Cela nous a permis de récolter 136 textes, soit plus de la moitié de notre corpus, contenant de nombreuses erreurs, mais nous avons introduit des biais dans l'acquisition de ces données, ne serait-ce que par la situation de scription qui est peu naturelle. La dictée n'est en effet généralement pas un exercice réalisé sur ordinateur.

Tout d'abord, nous avons sélectionné les textes à dicter en fonction de leur contenu grammatical et des erreurs qui pouvaient être commises. Nous avons donc ainsi sans doute induit des erreurs. Par ailleurs, les dictées ont été réalisées dans un navigateur internet, certains ayant la fonction de correction orthographique activée. Même si ce sont

principalement les erreurs grammaticales qui nous intéressent, cela introduit une disparité entre les textes dont les scripteurs ont pu corriger l'orthographe et les autres. Par ailleurs, ce sont des personnes tierces qui ont dicté les textes aux étudiants, et nous n'avons donc pas eu de contrôle sur la manière dont ces dictées ont été réalisées.

En forçant le rédaction de textes, nous avons sans doute influé sur les erreurs commises. Il nous faudra donc en tenir compte lors de l'analyse du corpus et ne pas généraliser ses erreurs à tous les scripteurs.

4 CONSTITUTION DU CORPUS

En dépit des contraintes qui se sont imposées, nous avons essayé de diversifier nos textes au maximum, tant au niveau des types de documents qu'au niveau des scripteurs et des situations de scription. Nous n'avons cependant pu recueillir que 4 types de documents. Nous avons annoté leurs erreurs une par une à l'aide d'une typologie existante que nous avons adaptée à notre corpus, à la suite de quoi nous avons pu faire quelques premières observations sur les erreurs collectées.

4.1 Recueil des données

4.1.1 MÉTHODOLOGIE DE RECUEIL DES DOCUMENTS

Nous avons recueilli 4 types de documents selon des protocoles différents.

- Des dictées : Nous avons sélectionné sur Internet 10 courts textes choisis soit pour leur longueur soit parce qu'ils comportaient quelque difficulté grammaticale. Ces textes ont été donnés en dictée à 136 étudiants de L1 à L3. Les dictées ont été réalisées sur une page Internet que nous avions développée, permettant d'une part d'éviter l'utilisation d'un correcteur grammatical, d'autre part d'enregistrer automatiquement la dictée dans une base de données ainsi que des informations sur les scripteurs (âge, sexe, niveau d'étude et langue maternelle). Le principe de la dictée, avec une contrainte temporelle forte, nous a semblé idéal pour recueillir un grand nombre d'erreurs de grammaire. Il a en effet été montré qu'un temps limité pour la situation de scription est un facteur augmentant les erreurs (Lucci et Millet, 1994).

- Des résumés : Nous avons également intégré dans notre corpus 36 résumés réalisés par des étudiants du M1 au doctorat, au cours d'une expérience de recherche d'un projet européen, à partir de textes scientifiques longs. Le degré de formalité était plutôt faible, les résumés étant en effet destinés à être traités automatiquement et non pas à être relus par un humain. Les variations sur la quantité d'erreurs observées sont en grande partie dues à la durée variable passée par chaque sujet à rédiger les résumés.

- Des emails : Pour diversifier les types de scripteurs, nous avons recueilli 42 emails de la liste de diffusion des utilisateurs d'OpenOffice.org. Nous avons passé en revue les textes de 16 personnes non étudiantes nous ayant donné leur accord, et nous avons gardé ceux contenant au moins une erreur de grammaire ou d'orthographe. S'agissant d'emails, il n'y avait aucune contrainte de temps pour les scripteurs. Par contre, le degré de formalité était très variable, les emails rédigés le moins formellement contenant le plus d'erreurs.

- Des textes de FLE : Afin que la population de scripteurs non francophones natifs soit représentée dans notre corpus, nous avons également récupéré sur deux blogs d'enseignantes de français langue étrangère (FLE) 14 commentaires rédigés en français par des étudiants étrangers.

4.1.2 STOCKAGE DES DONNÉES

Afin de stocker facilement les documents du corpus, nous avons réalisé une interface en PHP permettant d'enregistrer chaque nouveau texte sur le serveur et de lui adjoindre toutes les

informations disponibles sur son type, sa date d'acquisition ,sa source, sur le scripteur (âge, sexe, niveau d'étude et langue maternelle) et sur la situation de scription (formalité et contrainte de temps). Ces informations sont enregistrées en base de données.

Cette base de données reçoit également par la suite les informations précises sur chaque erreur contenue dans le document et qui a été annotée manuellement en XML. Une table rassemble toutes les erreurs avec pour chacune son ID, les ID du document auquel elle appartient et de la phrase dans laquelle elle se trouve, la phrase complète, la phrase en version annotée, le type de l'erreur, ce qui était attendu, le moyen de détecter l'erreur, le texte erroné et la correction.

L'enregistrement de toutes ces informations en base de données nous permettra par la suite d'effectuer des recherches ou des analyses de manière aisée.

4.2 Annotation du corpus

4.2.1 FORMALISME D'ANNOTATION

À l'aide d'un petit script en PHP, nous avons extrait les données de leur base et les avons transcrites dans le formalismes XML afin de pouvoir les traiter facilement. Nous avons alors entrepris l'étiquetage manuel en XML des erreurs contenues dans les documents, à l'aide d'un logiciel spécialisé dans l'édition de fichiers de ce type.

Chaque erreur appartient à une phrase d'indice *idp* comprise dans les balises `<Phrase idp="ID26"> ... </Phrases>`. L'erreur elle-même est délimitée par les balises `<erreur>...</erreur>`. La balise `<erreur>` contient divers attributs permettant de caractériser l'erreur : son numéro *ide*, son type *ling* (type linguistique de l'erreur), ce qui était attendu *att*, une explication sur la manière d'identifier l'erreur *expl*, ainsi parfois que l'identifiant de l'élément permettant la détection de l'erreur *idref*.

Dans la balise `<erreur>` se trouve deux balises : `<Initial>...</Initial>` qui contient le mot tel qu'écrit au départ, et `<Correction>...</Correction>` qui contient la version corrigée.

Ainsi, une phrase est étiquetée de la manière suivante :

<Phrase idp="ID46"**>**
<REF idref="ID47"**>**Les divers formats**</REF>** de caractères sont
<erreur att1="sing pluriel" **expl1**="unif" **ide1**="ID48" **idref1**="ID47" **ling1**="ACC-pp-etre"**>**
<Initial>conservé**</Initial>**
<Correction>conservés**</Correction>**
</erreur> avant fermeture et après ouverture du document.
</Phrase>

4.2.2 DÉFINITION D'UNE TYPOLOGIE D'ERREURS

Pour effectuer l'étiquetage des erreurs, il nous fallait disposer au préalable d'une typologie. Or, un des buts de notre corpus est justement de nous permettre de définir cette typologie. Nous nous sommes donc inspirée de la typologie d'erreurs élaborée par (Granger, 2007) pour son corpus d'apprenants FRIDA afin de pouvoir commencer l'annotation. Nous l'avons adaptée en commençant par supprimer certains types qui ne nous semblaient pas pertinents concernant la correction grammaticale automatique, telles les catégories d'erreurs se référant à des problèmes d'ordre sémantique ou stylistique. Puis nous l'avons complétée au fur et à mesure de l'avancée de l'étiquetage et de la découverte de nouvelles sortes d'erreur. Nous avons ainsi ajouté des catégories à cette typologie, ou plutôt nous l'avons affinée. Nous avons par exemple plusieurs textes dans lesquels toutes les apostrophes ont été omises. Nous avons donc

créé le type 'Orth-apost' spécifiquement pour ce type d'erreur particulier aux textes dactylographiés. Dans certains textes, souvent les mêmes que ceux que nous venons d'évoquer, tout signe diacritique a également été omis volontairement, pour simplifier et accélérer la vitesse de saisie. Nous ne l'avons pas encore fait, mais il pourrait être pertinent de distinguer les omissions volontaires systématiques des accents, des erreurs d'accentuation classiques et isolées.

Pour certains textes, de FLE principalement, nous avons également dédoublé le type défini par (Granger, 2007) pour les mots redondants. Nous avons en effet distingué d'une part les répétitions d'un même mot *("...après trois ans **passés passés** en cure..."),* et d'autre part les ajouts de mots ("*D'abord, il neigeait **le** vendredi dernier*").

Nous avons finalement établi 25 types d'erreurs que nous avons répartis dans 8 catégories.

60. Acc-SN : erreurs d'accord sur les éléments des syntagmes nominaux (Noms, adjectifs et déterminants)
61. Acc-PP : erreurs d'accord sur les participes passés, après l'auxiliaire être aussi bien qu'avoir.
62. Rplct : lorsqu'un mot est remplacé par un autre par confusion ou homonymie, ou lorsqu'un phénomène d'euphonie est mal appliqué.
63. Orth : erreurs d'ordre orthographique, c'est-à-dire qui génèrent des mots inconnus (abréviation, accentuation, agglutination, segmentation, apostrophe, graphie, morphologie et typographie)
64. Verb : erreurs concernant les verbes (erreurs de mode, de temps, de personne ou d'auxiliaire)
65. Ponct : erreurs de ponctuation
66. Synt : erreurs d'ordre syntaxique, à savoir l'oubli, l'ajout ou la répétition d'un mot.
67. Maj : lorsqu'il manque une majuscule, en début de phrase principalement

4.3 Premières analyses et résultats

4.3.1 ERREURS OBSERVÉES

À partir des 237 textes de notre corpus, nous avons extrait 738 phrases erronées avec en tout 1991 erreurs, soit en moyenne 1,5 erreur par phrase en général, et presque trois par phrase erronée. 53% de ces erreurs, soit un peu plus de la moitié, sont des erreurs d'orthographe, le reste des erreurs étant d'ordre grammatical, c'est-à-dire non détectable par un correcteur orthographique qui se contente de vérifier l'existence d'un mot dans un lexique.

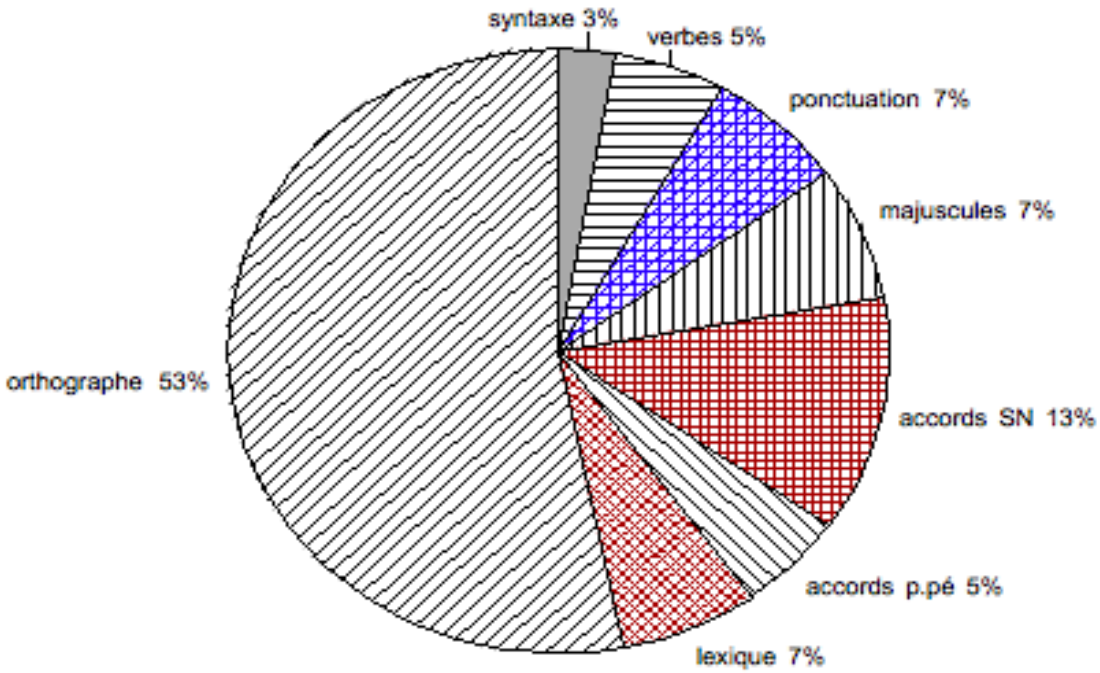


Figure 4 : Répartition des erreurs par catégories

Les erreurs grammaticales les plus fréquentes concernent les accords au sein des syntagmes nominaux (13%), majoritairement entre les déterminants et les noms. Viennent ensuite en proportions égales (7%) l'omission de la majuscule, principalement en début de phrase, les erreurs de lexique dues le plus souvent à l'homonymie, et les erreurs de ponctuation. Les accords des participes passés après l'auxiliaire être ou l'auxiliaire avoir lorsque l'objet est antéposé, et les erreurs sur les verbes qui sont le plus souvent des problèmes d'accord avec le sujet représentent chacun 5% des erreurs du corpus. 3% seulement des erreurs concernent la syntaxe

4.3.2 PERSPECTIVES D'ANALYSES

Le fait d'avoir stocké en base de données toutes les informations sur chaque erreur va nous permettre d'effectuer facilement de nombreuses analyses des erreurs en fonction des divers paramètres dont nous disposons (type de scripteur, situation de scription, etc.). Une interface réalisée en PHP nous permettra d'interroger la base afin notamment d'éditer des graphiques pour des données statistiques que nous pourrons comparer avec des données de textes manuscrits, ou bien de visualiser toutes les erreurs d'un type donné pour analyser leurs contextes et mieux gérer leur détection dans notre outil.

5 CONCLUSION

Pour résoudre les problèmes de la correction grammaticale automatique et concevoir un nouvel outil, nous avons choisi une approche corpus nous permettant d'analyser les erreurs des dactylographes. Tout en sachant que nous ne pourrions pas atteindre la représentativité des documents dactylographiés, nous avons constitué un corpus d'erreurs à partir de divers textes tapuscrits. Nous avons repris une typologie d'erreurs que nous avons adaptée à nos besoins et nous avons ainsi pu annoter tout le corpus. Nous avons stocké toutes les données en base de données afin de pouvoir facilement, à l'aide d'une interface que nous allons développer, réaliser des analyses diverses du corpus. Pour le moment nous avons pu identifier les principales erreurs du dactylographe. L'orthographe, largement majoritaire, devra

impérativement être traitée par le correcteur grammatical que nous allons développer. Nous devrons aussi nous concentrer sur les erreurs d'accords nominaux et verbaux et sur les confusions d'homophones.

Dans un second temps, notre corpus servira également à tester notre outil. Il constitue en effet un bon échantillon des erreurs les plus courantes que notre outil devra être capable de détecter.

6 RÉFÉRENCES

Abeillé A. (1993). *Les nouvelles syntaxes. Grammaires d'unification et analyse du français.* Armand Colin.

Abney S. (1991). « Parsing by chunks ». Dans R. C. Berwick, S. Abney et C. Tenny, *Principle-Based Parsing : Computation and Psycholinguistics*. Boston : Kluwer Academic Publishers. p. 257-278

Boissière P. et al. (2007). « Méthodologie d'annotation des erreurs en production écrite. Principes et résultats préliminaires ». *Actes de TALN 2007.*

Bonnard H. (1981). « Code du français courant ». Éditions Magnard.

Equipe COVAREC (1994). *Corpus de variations orthographiques.* Laboratoire LIDILEM, Université Stendhal-Grenoble 3.

Granger S. (2007). « Corpus d'apprenants, annotation d'erreurs et ALAO : une synergie prometteuse ». *Cahier de lexicologie*, 2007-2. .

Grevisse M. (1993). *Le Bon Usage.* Treizième édition par André Goosse. Éditions De Boeck – Duculot.

Habert B. et al. (1997). *Les linguistiques de corpus*. U Linguistique, A. Colin.

Lebarbé T. (2002). *Hiérarchie inclusive des unités linguistiques en analyse syntaxique coopérative; le segment, unité intermédiaire entre chunk et phrase dans le traitement linguistique par système multi-agents.* Thèse de Doctorat, Université de Caen.

Lechelt M. (2005). *Analyse et conception d'un correcteur grammatical libre pour le français*. Mémoire de stage, Master 2 Industries de la Langue, Université Stendhal-Grenoble 3.

Lucci V. et Millet A. (1994). *L'orthographe de tous les jours, enquête sur les pratiques orthographiques des français.* Éditions Champion.

Sinclair J. (1996). *Preliminary recommendations on corpus typology.* Document technique, Eagles.

Souque A. (2008). « Vers une nouvelle approche de la correction grammaticale automatique ». *Actes de RECITAL 2008*.

Tesnières L. (1959). *Éléments de syntaxe structurale*. Klincksieck.

L'ANALYSE MULTIVARIÉE DES PRODUCTIONS VERBALES DE JEUNES ENFANTS :UN ÉLÉMENT DE PLUS EN FAVEUR DES CORPUS LONGITUDINAUX

Frédéric Torterat
Université de Nice / IUFM
frederic.torterat@unice.fr

1 INTRODUCTION

Les verbalisations des jeunes enfants, en particulier dans un cadre dialogal, témoignent d'acquisitions diverses. La prise en compte de cette diversité, par les personnes qui encadrent les enfants et qui les accompagnent, suppose d'en retenir plusieurs domaines de variabilité, laquelle est interindividuelle en ceci que les acquisitions ne sont pas les mêmes d'un enfant à l'autre, mais aussi intra-individuelle, en ceci que les capacités verbales d'un même enfant sont l'objet de constantes modifications. A cet égard, si l'analyse du premier domaine de variabilité peut s'appuyer sur ce que nous nommerons des corpus instantanés, celle du deuxième implique la plupart du temps des corpus longitudinaux (plusieurs instantanés d'un moment M1 à un moment Mn, et pour les mêmes enfants).
D'une manière assez générale, l'approche de ces types de corpus ne s'avère productive que dans la mesure où elle aboutit à des éléments révélateurs, que nous appellerons, comme cela se pratique par ailleurs, significatifs en acquisition. Quand ils sont convertis en données proportionnelles, les éléments discursifs appellent plusieurs formes de traitement, dont il est quelquefois difficile de prévoir les profits pour la pédagogie ou la recherche (Sinclair 2005, Pincemin 2007). Dans cette vue, l'une des possibilités qu'apporte aujourd'hui le traitement informatique des données, à savoir, pour ce qui nous occupe ici, l'analyse multivariée, permet d'établir des liaisons statistiques entre plusieurs variables, et de rendre compte, quand la démarche présente un minimum de garanties méthodologiques, d'un système de relations à l'appui desquelles on peut envisager de désigner des caractéristiques communes d'une part, et d'autre part distinctes, autrement dit, plus simplement, des rapprochements et des écarts.

Couramment employé en sociologie, et de plus en plus en linguistique de l'acquisition, ce type de traitement, qui est *multivarié* en ceci qu'il implique des composantes multiples, confirme par exemple l'existence de groupes sociaux, d'habitudes culturelles ou de comportements interindividuels parmi lesquels se dégagent éventuellement des tendances (Cibois 2007). Il en est ainsi de pratiques qui présentent simultanément les mêmes caractéristiques, ou de groupes d'individus qui ont un comportement analogue : autant d'éléments qu'une représentation unifiée permet d'aborder dans un ensemble où s'organisent des liaisons plus ou moins effectives que les statisticiens, à la suite des travaux de différents mathématiciens (voir Martin 1997), désignent à travers des attractions et des répulsions entre un nombre plus ou moins élevé de variables.

Rappelons toutefois que les résultats d'une analyse multivariée (désormais AM) ne répondent à « aucune métrique simple », pour reprendre l'expression de Rastier (2009), et qu'ils apportent peu d'informations s'ils ne rejoignent pas assidûment les données du corpus, et donc le corpus lui-même. Qui plus est, pour peu que les variables renvoient à des données

brutes ayant des liens discutables avec l'objet de l'analyse, un tel traitement informatique a toutes les chances de présenter un certain effritement. Enfin, dans le cas où des regroupements s'organisent entre des variables renvoyant à des allants-de-soi particulièrement improductifs, les conclusions intermédiaires peuvent déboucher sur des facteurs triviaux, qui sont d'un gain minime.

D'autres défauts peuvent apparaître au cours du traitement, et c'est donc suivant la minutie avec laquelle s'effectue l'approche méthodologique des données brutes que ces deux objets de mésestime – dispersion de la recherche et trivialité des composantes représentées – seront pris en compte ou non.

Car les atouts d'une AM ne sont pas négligeables : d'une part, ce type d'analyse décrit une multiplicité à partir de laquelle elle dégage des regroupements ; d'autre part, elle conduit à concrétiser, à partir d'un corpus converti par exemple en plusieurs (sous-)ensembles de proportions, des tendances quantifiables à travers ce qu'on appelle la « contribution aux facteurs », avec des domaines de variance différents selon les combinaisons de variables opérées.

Le corpus présenté dans ces pages n'est pas extrait d'une base de données généraliste de type CHILDES (Mac Whinney 2000) ou de données précédemment converties, mais porte sur les productions spontanées d'enfants enregistrés dans un contexte interlocutif de terrain, de manière à fournir des éléments de corpus oraux sur des bases accessibles pour des non statisticiens, et ce dans un temps court. Cette approche permet aux informations d'être employées spécifiquement dans le cadre de la professionnalisation des professeurs des écoles, d'autant que l'annotation des corpus oraux n'est pas sans poser des problèmes spécifiques (Benzitoun 2004), et que la multi-annotation exige une autre forme d'accompagnement. Collecté auprès de classes de maternelle à Nice (en circonscription), le corpus en question, intitulé DAM07, est constitué de productions correspondant à trois « dialogues de classes » encadrés, dans les mêmes demi-journées et sur des supports pédagogiques analogues, par M (professeur des écoles à Nice en 2007). Les effectifs « nominatifs » renvoient donc à des groupes restreints de petites et moyennes sections de maternelle, avec dans notre cas des enfants âgés de 2,5 à 4,2 ans, et des matériaux qui représentent un « petit » ensemble (expérimental) de 12000 mots.

Qu'on nous permette d'indiquer que les objectifs professionnels de cette recherche ont d'abord été praxéologiques : il s'agissait de déterminer comment organiser les groupes de manière à prendre en compte la diversité des acquisitions verbales des enfants, avec comme possibilité de mesurer les (ré)emplois des marqueurs de structuration discursive à travers l'analyse de cette diversité. Les données – hors caractéristiques phonologiques – ont donc suscité plusieurs démarches, dont nous reportons ci-après le suivi et les premiers résultats.

2 APPROCHE MÉTHODOLOGIQUE

Le corpus discursif sur lequel nous nous appuyons renvoie à des productions verbales qui, au moment des retranscriptions, ont été faiblement redressées. Etant donné que les données acoustiques et prosodiques n'ont pas été prises en compte, la ponctuation qui a été employée dans le format « texte », pour ce qui la concerne, n'apporte de son côté aucun recours explicatif ou descriptif. Pour autant, les éléments collectés nous ont parus suffisamment robustes pour l'analyse : sans doute n'est-ce pas la « grandeur » du corpus qui garantit sa représentativité (à moins de ne l'envisager qu'à travers les chaînes de caractères), mais sa consistance pour l'analyse. Pour celle-ci, nous avons envisagé la diversité des productions discursives bien sûr, mais aussi les variables qu'elles sont en mesure d'impliquer, les regroupements qu'elles permettent d'envisager, ainsi que leur apport d'informations non

préétablies. Cette démarche, dans l'ensemble, s'est alors assigné pour objets de dégager des données quantitatives qui, dans un deuxième temps, ont été confrontées à des éléments caractérisés de manière plus aboutie.
En voici tout de suite un extrait pour illustration, lequel est reporté en « texte linéaire » tel qu'il a été retranscrit une première fois, sans les annotations ni les remplacements correctifs (l'abréviation « M » renvoie à l'intervenante) :

> G1 : Les coqs (enfants grands à moyens parleurs). Sont présents : Marie-Sarah, Léo, Noémie, Mattéo B, Apolline, Mathilde.
>
> Léo : Qu'est-ce que tu vas enregistrer ?
>
> M : Je vais vous enregistrer pendant que vous fabriquez les voitures. Alors, dites-moi, vous avez trié ce qui roulait ?
>
> Noémie : Moi c'est dans la boîte !
>
> Marie-Sarah : Moi aussi c'est dans la boîte.
>
> M : Pour fabriquer une voiture, qu'est-ce qu'on doit prendre ?
>
> Mattéo : Des roues !
>
> M : On doit prendre des roues...
>
> Léo : Et aussi la carrosserie.
>
> M : La carrosserie aussi, oui tu te souviens bien Léo. Alors choisissez tous une boîte pour faire la carrosserie de votre voiture.
>
> Mathilde : Moi je prends cette boîte.
>
> Noémie : Et moi cette boîte.
>
> M : D'accord. Comment vous allez vous y prendre pour accrocher les roues maintenant ?
>
> Noémie : Ben on va faire un petit trou. (...)

Les aménagements graphiques que nous avons pratiqués sur le corpus linéaire sont restés sommaires, d'autant que le moindre ajustement réclamant une justification (Dister et Simon 2008), nous avons tâché de ne pas compromettre le recours aux documents par des explications trop abondantes. D'autre part, nous avons reporté dans les premières transcriptions une ponctuation conforme à une présentation « académique » des verbatims, telle qu'elle apparaît dans de nombreuses monographies, compte tenu du fait, notamment, qu'il n'existe pas de correspondances régulières entre les phénomènes prosodiques, comme les contours intonatifs et les pauses, et la ponctuation graphique (Grobet 1997, Shriberg *et al*. 2000).

Dans la mesure, d'autre part, où les enregistrements ont été effectués dans de bonnes conditions matérielles, nous avons mis momentanément à l'écart de cette première approche les multitranscriptions, les disfluences répétitives, les commentaires para-verbaux et les cas de chevauchements. Sur le plan méthodologique, cela revient à dire que la réflexion menée a moins porté sur le traitement d'un corpus oral spécifique, que sur la mise en œuvre d'une démarche analytique.

Les productions verbales ont été intégrées à la base de données du logiciel Lexica (5.0), diffusé par *Sphinx Développement*, et acquis à la suite d'un appel d'offre de l'Université de Nice à travers sa composante en charge de la formation des enseignants. Comme d'autres logiciels de même type, Lexica rassemble des programmes de lexicométrie, de statistique textuelle et d'analyse multivariée qui facilitent en partie la détermination de normes de dépouillement déjà pratiquées par ailleurs en linguistique de l'acquisition (Cf. Sansonetti 2003). Etant donné que la conduite méthodologique engagée, dans ses grandes lignes, a d'abord consisté à dégager les domaines de variabilité interindividuelle entre les enfants, nous

avons soustrait M non pas du corpus linéaire, mais des éléments triés en données brutes. C'est donc à partir du corpus non linéaire (trié par ordre alphabétique balisé) et ouvert sous la forme d'un fichier de données « L » (*texte à analyser* sous Lexica), que le format texte précédemment balisé est passé en données ASCII.

Dans l'application pratique, au moment de confirmer le choix automatique des variables (les intervenants d'un côté, et les données textuelles de l'autre), on requiert des options maximales pour un « texte », en paramétrant peu à peu les balises[67]. Les variables apparaissent alors en « ouverte texte » (une *modalité* qu'il est possible de modifier par la suite). Une fois un premier dépouillement effectué sur l'instantané qui nous a servi de référence, les variables donnent les répartitions suivantes, lesquelles en représentent un premier abord.

Pour le Groupe 1 :

---	*Nombre de mots*	*Nombre moyen de mots*	*Nombre de mots différents*	*Nombre de mots uniques*	*Fréquence maximum*	*Mot le plus fréquent*
Apolline	40	4,44	12	11	2	roues
Léo	99	4,30	34	25	4	faut
Marie-Sarah	51	4,25	12	7	5	deux
Mathilde	51	4,25	21	18	2	roule
Mattéo	121	5,26	34	19	5	faire
Noémie	158	5,64	43	31	4	roule [68]

Pour le Groupe 2 :

---	*Nombre de mots*	*Nombre moyen de mots*	*Nombre de mots différents*	*Nombre de mots uniques*	*Fréquence maximum*	*Mot le plus fréquent*
Alain	22	3,67	10	6	2	Gaad
Ambre	105	4,20	31	20	4	Euh
Antony	12	3,00	6	5	2	deux
Kyllian	54	3,86	21	16	4	ça
Lucas	55	5,00	21	18	2	prends
Maxim	98	5,16	29	21	4	roues
Ophélio	39	2,79	18	16	5	roues [69]

[67] Si ce n'est : « l'ordre des balises sépare les observations » / « les balises sont numérotées et des parties créées ».

[68] Mots les plus fréquents :
Apolline : roues (2) ; petite (1) ; roule (1) ; arrive (1) ; veux (1) ; sais (1) ; faire (1) ; oui (1) ; grande (1) ; baguette (1) ;
Léo : faut (4) ; quatre (3) ; faire (3) ; regarde (3) ; ouais (2) ; roule (2) ; voiture (2) ; Julie (2) ; oui (2) ; Ah (1) ;
Marie-Sarah : deux (5) ; ça (2) ; veux (2) ; besoin (2) ; roues (2) ; Ah (1) ; oui (1) ; belle (1) ; mets (1) ; ici (1) ;
Mathilde : roule (2) ; boite (2) ; oui (2) ; ça (1) ; débloque (1) ; ici (1) ; sais (1) ; biscuits (1) ; roue (1) ; marche (1) ;
Mattéo : faire (5) ; roues (4) ; ça (3) ; tchou (2) ; fait (2) ; vais (2) ; mettre (2) ; quatre (2) ; trou (2) ; arrive (2) ;
Noémie : roule (4) ; faire (4) ; ben (3) ; ça (3) ; non (3) ; oui (3) ; trou (2) ; boîte (2) ; roues (2) ; réussi (2)

[69] Mots les plus fréquents :
Alain : Gaad (2) ; fait (2) ; trou (2) ; veux (2) ; ça (1) ; roule (1) ; peux (1) ; aider (1) ; roue (1) ; faire (1) ;
Ambre : Euh (4) ; maîtresse (3) ; trous (3) ; roule (3) ; roues (2) ; fait (2) ; grande (2) ; non (2) ; veux (2) ;

Pour le Groupe 3 :

---	*Nombre de mots*	*Nombre moyen de mots*	*Nombre de mots différents*	*Nombre de mots uniques*	*Fréquence maximum*	*Mot le plus fréquent*
Alain	28	2,33	11	7	3	Heu
Jaed	11	1,83	9	8	2	Ah
Mathis	8	2,67	5	4	2	veux
Sarah	23	2,88	11	7	3	oui
Shon	7	1,00	5	3	2	Ata[70]

Ces données ne sont pas, bien entendu, à mettre sur un même plan. Si les deux premières variables renvoient à des données strictement quantitatives, la troisième coïncide indirectement avec la diversification des productions, et la suivante avec celle de possibles acquisitions lexicales. Les résultats 4 à 6 représentent pour leur part un apport difficile à mesurer, d'autant que, pour ce qui relève du nombre de mots uniques par exemple, d'éventuelles faiblesses quantitatives peuvent en revanche témoigner, chez l'enfant, d'une certaine démarche interlocutive. Les trois derniers types d'informations (*mots uniques, fréquence, mot le plus fréquent*) sont par ailleurs directement liées aux exigences pédagogiques à l'oeuvre au moment du relevé, et varient suivant la pratique évaluative éventuellement mise en place, les consignes de l'exercice et les conditions de leur passation. D'autre part, à la faveur d'une première confrontation des données, il est apparu que le « nombre moyen de mots » (par intervention) n'apporte en général qu'une corroboration quantitative des données qui l'encadrent. Ces éléments ont donc été, au moins pour un temps, considérés comme intermédiaires. Il en a été de même pour les répartitions par *catégories*, qui, sur corpus en données brutes, spécifient la liste des 10 mots marqués du lexique (avec les nombres d'occurrences pour chaque mot), ici pour *Apolline* :

pas	4	ai	1	arrive	1	est	1
faire	1	peuvent	1	roule	1	rouler	1
sais	1	veux	1				

regarde (2) ;
Antony : deux (2) ; bâtons (1) ; ici (1) ; roues (1) ; non (1) ; côtés (1) ;
Kyllian : ça (4) ; marche (4) ; Regarde (4) ; Paw (2) ; fait (2) ; beau (1) ; Bah (1) ; fais (1) ; bâtons (1) ; deux (1) ;

Lucas : prends (2) ; brindilles (2) ; oui (2) ; cassé (1) ; drôle (1) ; euh (1) ; boîte (1) ; mer (1) ; montagne (1) ; dirait (1) ;
Maxim : roues (4) ; trous (4) ; faire (3) ; fais (3) ; fait (2) ; Kyllian (2) ; oui (2) ; Quatre (2) ; ça (1) ; Bah (1) ;
Ophélio : roues (5) ; oui (2) ; accrocher (1) ; ça (1) ; deux (1) ; après (1) ; mets (1) ; Heu (1) ; Quatre (1) ; manque (1)

[70] Mots les plus fréquents :
Alain : Heu (3) ; Julie (3) ; ça (2) ; roule (2) ; Aye (1) ; mal (1) ; veux (1) ; arrive (1) ; regarde (1) ; roue (1) ;
Jaed : Ah (2) ; oui (1) ; Iiou (1) ; veux (1) ; ça (1) ; mou (1) ; Yo (1) ; ya (1) ; bou (1) ;
Mathis : veux (2) ; maison (1) ; oui (1) ; pique (1) ; Papa (1) ;
Sarah : oui (3) ; deux (2) ; faut (2) ; Julie (2) ; belle (1) ; euh (1) ; roues (1) ; ça (1) ; accord (1) ; regarde (1) ;
Shon : Ata (2) ; Voiture (2) ; Heu (1) ; oui (1) ; roule (1)

elle est plus petite
elle / elle roule
et moi j'ai pas de roues
j'y arrive pas
moi aussi je veux
moi je sais pas le faire
oui celle là
une grande baguette
tes roues elles peuvent pas rouler ?[71]

Même si certains de ces éléments permettent d'envisager des regroupements, de manière à réordonner dans Lexica les données brutes en tableur, ils ne procurent qu'une visibilité réduite des domaines de variabilité qui nous occupent. Indiquons en revanche qu'un report des mêmes données concernant l'intervenante principale (« M »), montre de son côté dans quelles proportions l'adulte référent est effectivement intervenu parmi les verbalisations des enfants :

---	*Nombre de mots*	*Nombre moyen de mots*	*Nombre de mots différents*	*Nombre de mots uniques*	*Fréquence maximum*	*Mot le plus fréquent*
Groupe 1	928	13,45	178	112	15	roues
Groupe 2	889	12,01	149	88	22	faire
Groupe 3	610	14,52	139	99	9	ça

Ce dépouillement liminaire permet toutefois de noter les premières récurrences vers un « tableau à plat généralisé »[72], qui, pour le groupe 1 (*Marie-Sarah ; Mathilde ; Mattéo ; Noémie ; Apolline ; Léo*), apporte des indications non négligeables, comme le font apparaître les extraits suivants :

[71] Sur corpus lemmatisé, avec la liste des 9 mots marqués du lexique (et les nombres d'occurrences pour chaque mot), cela donne les données ci-après :

pas	4	rouler	2	arriver	1	avoir	1
être	1	faire	1	pouvoir	1	savoir	1
vouloir	1						

il être plus petit
il / il rouler
et moi j avoir pas de roue
j y arriver pas
moi aussi je vouloir
moi je savoir pas le faire
oui celui là
un grand baguette
ton roue il pouvoir pas rouler

[72]La manipulation informatique revient à se rendre à l'onglet « analyser », puis à cliquer sur « déterminer l'analyse » en indiquant les premières variables (ensemble des participants excepté M). Nous remercions chaleureusement Jean-Jacques Legendre pour ses explications sur tous ces points.

verbalisations
et aussi la carrosserie
et bien moi elle roule ma voiture
et c'est quoi ça / cet outil // c'est comme un bouchon
et moi aussi là / partout
et moi cette boîte
et moi j'ai pas de roues
moi aussi c'est dans la boîte
moi aussi je veux
moi ça me fait même pas mal
moi ça y est
moi c'est dans la boîte
moi c'est la boîte d'oeufs // ça c'est une boîte d'oeufs
moi ici
moi je l'ai fait
moi je peux pas / regarde
moi je prends cette boîte
moi je sais pas le faire
moi je vais les faire
moi je veux la mienne là
moi j'ai réussi à le faire
moi j'ai réussi à les faire
moi j'ai une roue qui marche pas très bien
moi j'arrive pas à le faire
moi tu m'as donné un grand // fais voir ton bâton
moi / là / là

En marge du caractère hasardeux et accumulatif de ces reports de productions non nominatives, ces énumérations nous conduisent à admettre la difficulté d'opérer des regroupements à partir de certains items. En outre, ces supports dénoncent le fait que les caractéristiques paradigmatiques, mais aussi syntagmatiques et topologiques, peuvent toutes avoir une capacité classificatoire. Cela suppose que nous écartions un relevé des « segments répétés » au profit d'un relevé d'éléments « récurrents », pour le traitement informatique desquels une identification manuelle s'impose.

Un tel tableau à plat généralisé valorise en partie, dans le même temps, la diversité des éléments du corpus, d'une part, et une possible représentativité des productions verbales, à partir desquelles il est intéressant d'établir des concordances spécifiques de manière à saisir la singularité de certaines données. Ainsi, des récurrences apparaissent quand l'opérateur interphrastique de transition *et* (*moi / aussi*) marque un lien paratactique entre le cotexte antérieur et les éléments qui lui sont postposés : tantôt thématisant (*et bien moi / et c'est quoi ça / et moi aussi là*), tantôt thématisant et rhématiseur (*et moi cette boîte*), il présente dans plusieurs cas une dimension cadrative, en ceci qu'il contribue à la structuration discursive du dialogue et des co-interventions des participants (Saussure et Sthioul 2002 *inter al.*). De même, quelques récurrences interviennent à l'appui du pronom *moi*, dont les multiples

combinaisons avec des déictiques (*ici, là, c'est..*) témoignent variablement de la prise en compte effective, par les enfants, du contexte interpersonnel. Ce monosyllabe est d'ailleurs suivi dans la plupart des cas d'un pronom dialogal (*je, je (là), me, tu*), et suppose dans presque tous un appariement au cadre interlocutif (les emplois de *moi* avec des éléments subséquents appartenant au cadre délocutif, comme dans *moi c'est dans la boîte / moi c'est la boîte d'oeufs*, sont moins répandus : cf. Apothéloz 1997).

C'est ici qu'intervient, en termes d'approche méthodologique des productions, la sélection de variables généralisables et l'exigence d'une caractérisation des éléments instanciés. En la matière, des entretiens menés à la suite d'analyses de pratiques auprès de professeurs des écoles (Nice, 11-2007 ; 02-2008) montrent que les premières démarches évaluatives que les intervenants mettent en œuvre pour *l'oral* prennent appui sur le nombre d'interventions, combiné éventuellement avec celui des mots et des items distincts. Plus marginalement, les « valeurs » que prennent les verbes fléchis, les syntagmes nominaux, les opérateurs et leur répartition peuvent intervenir, mais cette démarche analytique n'apparaît que chez 2 à 3 pour cent des personnes consultées, ce qui s'explique surtout par le fait des conditions de travail.

Comme il s'agit d'estimer les éventuelles proximités entre cette première analyse et une analyse linguistique plus aboutie, l'un des enjeux de la recherche a ainsi consisté à déterminer l'appropriété des valeurs pressenties, mais aussi, dans le même temps, l'opportunité même de tels corpus instantanés.

3 PREMIER TRAITEMENT DES DONNÉES

3.1 Pr®sentantion de la d®marche

A la suite d'une première analyse lexicométrique (« atelier lexical » dans Lexica), les regroupements de variables ont été effectués vers des tableaux croisés, avec les variables nominatives dans la première colonne (renvoyant aux enfants), et les variables descriptives dans la première ligne (renvoyant aux proportions de productions), à partir desquelles apparaissent ensuite les valeurs[73]. Le tableau croisé intègre ainsi un fichier en « données externes » (la première ligne indiquant les variables fermées), dont on demande une analyse factorielle qui correspond plus exactement à une ACP (analyse en composantes principales)[74]. L'ACP prend alors en compte les données descriptives et devient significative, à savoir que, pour ce traitement « 1 » et sur le même instantané de référence, ces données sont les suivantes : les interventions (spontanées et sollicitées), les mots en données brutes et les mots distincts (récurrents ou non), avec les proportions reportées ci-après.

Pour le Groupe 1 :

[73] Il suffit d'enregistrer l'ensemble sous la forme d'un tableau croisé en recodant les variables en « fermées multiples », pour reproduire ensuite les données, éventuellement en les transposant sous la forme d'un tableau xls.

[74]La manipulation consiste ici à se rendre sur l'onglet « analyses », puis sur « tableau croisé » : la boîte de dialogue permet ensuite de mettre la variable 1 en colonne et les autres variables « en ligne », en cliquant sur « regrouper dans un même tableau ».

	INTERV	MOTS	MOTS DIST
APOLLINE	9%	8%	7%
LEO	21%	19%	22%
MARIE SARAI	11%	10%	8%
MATHILDE	11%	10%	13%
MATTEO	22%	23%	22%
NOEMIE	26%	30%	28%

Pour le Groupe 2 :

	INTERV	MOTS	MOTS DIST
ALAIN	6%	6%	7%
AMBRE	27%	27%	22%
ANTONY	4%	3%	4%
KYLLIAN	15%	14%	15%
LUCAS	13%	15%	15%
MAXIM	20%	25%	21%
OPHELIO	15%	10%	16%

Pour le Groupe 3 :

	INTERV	MOTS	MOTS DIST
ALAIN	32%	36%	27%
JAED	17%	14%	22%
MATHIS	8%	10%	12%
SARAH	23%	31%	27%
SHON	20%	9%	12%

Les valeurs indiquées correspondent aux proportions individuelles de production chez les enfants. Ainsi Mattéo, dans le premier groupe, accapare-t-il 22 pour cent des interventions parmi l'ensemble de celles relevées chez les enfants, là où les productions de Sarah, par exemple, occupent 31 pour cent des mots produits par le groupe 3. Ces proportions s'inscrivent, comme nous l'avons précédemment indiqué, dans le format d'une analyse multivariée (plusieurs variables sont rassemblées). Or, bien qu'il existe des corrélations entre les trois types de proportions, celles-ci ne sont toutefois pas automatiques (Cf. les exemples d'Ophélio et de Shon, dans les groupes 2 et 3).

Dans tous les cas, l'une des premières questions à soumettre au traitement consiste à estimer dans quelle mesure une représentation, pour ainsi dire liminaire, de ces données, devient productive à partir de ces « variable(s) en ligne » et « variable(s) en colonne ». Ces résultats donnent effectivement des informations succinctes, en ce qu'ils ne font que confirmer des seuils plus ou moins élevés de dispersion du groupe, et déterminent en partie les premiers « écarts-type » entre les productions que les enseignants pressentent assez spontanément. Par ailleurs, ces premières données sont proprement cumulatives (interventions, mots et mots distincts), et ne concernent que les productions verbales en données brutes. En appliquant quoi qu'il en soit les variables en échelles[75], on peut obtenir un tableau de moyennes caractéristique[76] : l'analyse multivariée résume alors des tendances statistiques qui organisent des variables descriptives facilement convertibles.

L'ACP a pour mérite d'indiquer le positionnement des variables descriptives les unes par rapport aux autres (représentées ci-après sous forme de vecteurs). Pour le groupe 2 par

[75] Les variables « 1 » nominatives restent dans une fermée multiple, mais les autres variables (descriptives) passent de « fermées multiples » à « fermées échelles ».

[76] Ce qui revient à cliquer sur « analyser » dans la présentation du corpus, puis sur « tm », en croisant 1 et les variables descriptives regroupées.

exemple, on remarque que les écarts entre les vecteurs sont très faibles, alors que deux d'entre eux coïncident dans le cas du premier groupe (avec un écart sensible pour le troisième), et que l'ensemble des variables accusent des écarts importants, ce qui signifie que les données brutes ne sont pas *co-orientées*, comme le dégagent les contributions aux facteurs ci-après[77].

Pour le Groupe 1 :

	Axe 1 (+69.67%)		Axe 2 (+30.31%)	
CONTRIBUTIONS POSITIVES	INTERV	+46,0%	MOTS DIST	+92,0%
	MOTS	+46,0%		
CONTRIBUTIONS NEGATIVES			INTERV	-3,0%
			MOTS	-3,0%

Pour le Groupe 2 :

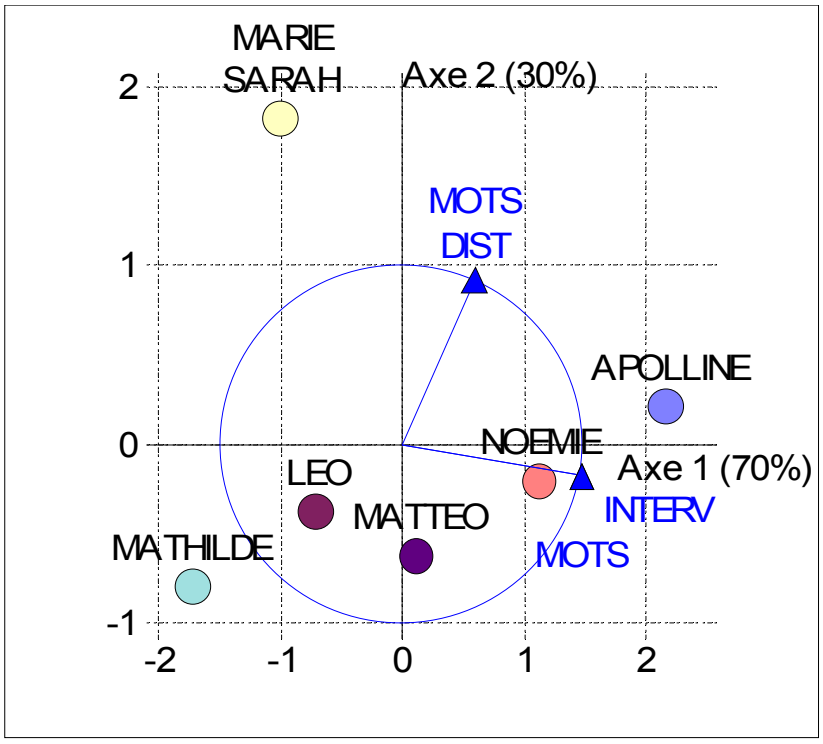
	Axe 1 (+95.40%)		Axe 2 (+4.03%)	
CONTRIBUTIONS POSITIVES	MOTS DIST	+34,0%	MOTS	+67,0%
	INTERV	+33,0%		
CONTRIBUTIONS NEGATIVES			INTERV	-18,0%
			MOTS DIST	-14,0%

Pour le Groupe 3 :

	Axe 1 (+44.63%)		Axe 2 (+33.29%)	
CONTRIBUTIONS POSITIVES	MOTS	+49,0%	MOTS DIST	+76,0%
	MOTS DIST	+11,0%	INTERV	+23,0%
CONTRIBUTIONS NEGATIVES	INTERV	-38,0%		

Plusieurs représentations graphiques existent pour figurer comment se positionnent les variables les unes par rapport aux autres, mais elles n'en sont ni un compte rendu exact, ni encore moins un aboutissement. Si nous prenons le groupe 1 par exemple, avec un taux de variance pour le moins satisfaisant (du fait surtout des faibles quantités de données), une représentation graphique simplifiée place les variables de la manière suivante :

[77] Dans le tableau de bord constitué, et après avoir recodé les variables descriptives en « fermées échelles », on clique sur « analyses » puis sur « tableau de moyennes / corrélation ». Apparaît une boîte de dialogue. On clique alors sur « pour chaque modalité / valeur de » en mettant « 1 », puis on sélectionne dans les « variables numériques » l'ensemble des variables échelles. Un « groupe 2 » apparaît dans le tableau de bord, en dessous duquel on requiert l'ACP.



L'approche de ce domaine de variabilité interindividuelle impliquant des proportions, elle s'appuie sur des valeurs numériques que les représentations graphiques, à travers l'ACP, projettent sur un plan bidimensionnel dont seuls les segments des vecteurs INTERV, MOTS et MOTS DIST, *a priori*, indiquent la tridimensionnalité (plus ils sont « courts », plus ils se rapprochent du centre). Toutes les valeurs sont ainsi résumées dans le graphique de manière à dégager des composantes principales, lesquelles rassemblent des ensembles de « points » corrélés. Les « contributions aux axes » nous indiquent alors dans quelle mesure telle ou telle variable prend part à la concrétisation des facteurs, laquelle paraît productive ici, vu que plus la somme des taux de variance se rapproche de 100, plus la représentation graphique a des chances d'« expliquer » les liens qui existent entre les variables. Dans le cas reporté ci-dessus, la somme des taux des facteurs 1 et 2 donne précisément 100, ce qui suppose que le graphique explique intégralement les relations qui s'établissent entre les valeurs.

Les tableaux de contributions et les représentations qui leur correspondent confortent quoi qu'il en soit le premier dépouillement pour ce qui relève des variables descriptives. En revanche, elles en facilitent l'abord en ceci qu'elles donnent un peu plus d'indications sur la manière dont se placent les variables nominatives. Or, non seulement il est apparu que les représentations présentent des taux de variance tout à fait satisfaisants sur l'ensemble des groupes, mais qu'elles permettent aussi d'esquisser les moyennes productives en acquisition, en mesurant les écarts quantitatifs et qualitatifs entre les interventions.

Indiquons qu'une autre possibilité, pour ce type d'analyse, est de déterminer les rapprochements éventuels entre les participants, lesquels rapprochements constituent l'un des critères de construction ou de déconstruction du groupe restreint : leur absence marquera éventuellement ce que nous appellerons un domaine de dispersion élevé, alors que plusieurs rapprochements au sein d'un même groupe peuvent marquer une réduction de la dispersion[78]

[78] Ces derniers éléments sont également envisageables en termes d'insertion, comme en témoigne l'exemple d' « Alain », qui est intégré à la fois dans le deuxième et dans le troisième groupes : dans le premier, il paraît peu inséré, alors que dans le deuxième, son insertion est d'autant plus effective que les écarts sont moindres par

(à condition bien sûr que ces rapprochements ou ces écarts soient validés par la matrice des corrélations et les coordonnées des variables). Sans doute cela apparaîtrait-il davantage avec une application des régressions, que nous ne reprendrons pas ici, mais les calculs de distance sont déjà, de notre point de vue, suffisamment adaptés à cette première démarche.

4 DEUXIEME TRAITEMENT

Afin d'établir dans quelle mesure les enfants opèrent effectivement une structuration discursive de leurs interventions, nous avons mis en oeuvre un deuxième traitement, auquel le premier a été confronté, et qui s'est effectué sur la base d'une analyse linguistique plus aboutie.

Sur les mêmes éléments du corpus de référence, les variables descriptives retenues cette fois-ci ont été les éléments verbaux (éventuellement passivés, négativés ou « amassés » : Gerdes et Kahane 2006), les éléments thématiques et rhématiques (au sens de Mel'čuk 2001), les éléments cadratifs (pour beaucoup déictiques : Charolles *et al.* 2005), et les opérateurs spécifiques (coordonnants et subordonnants en particulier : Torterat 2002, Rebuschi 2002, Desclés 2008), avec les abréviations respectives VERB, THEM, RHEM, CADR et OP. D'autres *témoins* (opérateurs évaluatifs par exemple) auraient pu être pris en compte pour l'étude, pour peu qu'ils eussent été à même de rendre compte d'un processus d'acquisition où leur présence est significative. Pour nous en tenir aux phénomènes de production et de structuration discursives, nous avons caractérisé le corpus non linéaire en nous conformant à l'approche précédente, sans passer par une multi-annotation.

L'autre type de réponse que ce deuxième traitement informatique est appelé à donner revient à définir si une analyse multivariée à l'appui de facteurs linguistiques plus aboutis, confirme en partie, ou non, le premier traitement établi en données brutes, dont les résultats peuvent présenter éventuellement une certaine trivialité, ou du moins se restreindre à la prise en compte de répartitions certes représentatives, mais pour le moins sommaires.

En pratique, la réédition du corpus (en données textuelles balisées) a consisté à étiqueter les éléments linguistiques un par un, en effectuant éventuellement des regroupements locaux, comme c'est le cas par exemple pour les amas verbaux et les syntagmes nominaux autour d'un même noyau prédicatif. Dans le format texte, les verbes et amas verbaux sont indiqués par commodité en italiques (avec éventuellement un trait en bas (*touche 8*) pour spécifier qu'un premier élément forme un tout avec un élément distant), les THEM sont indiqués en majuscules, les RHEM en petites majuscules, les CADR en souligné, et les OP en gras. Voyons ci-dessous trois extraits du corpus caractérisé :

Mathilde : ÇA *se débloque*
Mathilde : ICI
Mathilde : JE *roule /* (JE *roule*)
Mathilde : JE *sais pas*
Mathilde : LA BOITE DES BISCUITS
Mathilde : moi J'*ai* UNE ROUE **qui** *marche pas* TRÈS BIEN
Mathilde : moi JE *prends* CETTE BOÎTE
Mathilde : moi TU M'*as donné* UN GRAND // *fais voir* TON BÂTON
Mathilde : OUI JE *veux* UN TROU | LÀ
Mathilde : OUI

rapport aux autres, et qu'un domaine de connivence s'établit.

Mathilde : UN / DEUX
Mathilde : voilà

Lucas : *c'est* CASSÉ
Lucas : *c'est* DRÔLE
Lucas : *il y a* ~~euh~~ // UNE BOÎTE ?
Lucas : JE *prends* CELLE-LÀ
Lucas : JE *prends* LES BRINDILLES | À LA MER **et** LÀ | LES BRINDILLES | À LA MONTAGNE
Lucas : ON *dirait* UNE ÉPÉE // ~~Tayaya~~
Lucas : OUI *c'est* VRAI ÇA *marche*
Lucas : OUI
Lucas : *regarde* **comment** ELLE *est* ma voiture
Lucas : VOITURES
Lucas : ON *joue* À LA COURTE PAILLE ?

Shon : *ata*
Shon : *ata*
Shon : ~~heu~~
Shon : OUI
Shon : VOITURE
Shon : VOITURE
Shon : *roule*

La prise en compte par « regroupements » des éléments d'après Lexica ne permet pas un étiquetage automatique complètement satisfaisant, étant donné qu'il procède par assemblages pour une part fortuits et ne distingue pas certains (pseudo)homographes. De ce fait, nous avons opéré une réédition manuelle sur l'intégralité du corpus. Sans préconstruire en rien la quantification des données et comme indiqué *supra*, le relevé effectué nous a conduit à recouper des informations paradigmatiques, syntagmatiques et topologiques, ainsi que certaines informations contextuelles.

Plusieurs choix ont été effectués pour ce relevé, où l'on remarquera ci-après l'absence d'opérateurs spécifiques (comme les coordonnants et les subordonnants) dans le troisième groupe, et des écarts quelquefois importants, quoique prévisibles, entre les proportions des variables. Par exemple, les verbes et amas verbaux (où nous avons intégré les forclusifs de négation) s'accompagnent d'un *ata* de Shon qui, même si l'on peut estimer qu'il s'agit d'un cadratif grammaticalisé, renvoie à ce que nous avons identifié en contexte, momentanément du moins, comme un *attends* actionnel (Cf. Balthasar *et al.* 2003). D'autre part, il n'a pas été tenu compte, dans le relevé proportionnel, ni des opérateurs suspensifs du type *heu* (qui marquent des opérations diverses et nécessitent une prise en compte des contours prosodiques), ni des interjections plus ou moins onomatopéiques du type *ah*, *aye* ou *iliou* (lesquelles peuvent être plus ou moins thématiques, rhématiques ou cadratives : Wilkins 1992, Cuenca et Hilferty 1999).

Plusieurs occurrences du pronom *il* avec l'unipersonnel *falloir* apparaissent également dans les interventions, mais dans la mesure où, en marge de l'autocorrection qu'il suppose, ce pronom est non thématique sur le plan de la signification discursive, il a été intégré à l'amas

verbal dont il fait partie (cette sélection est bien évidemment discutable, mais les proportions qu'elle occupe dans les productions verbales sont minimes). Nous n'entrerons pas dans le détail des difficultés plus « fines » de cette caractérisation, qui fera l'objet d'autres contributions (Torterat 2010 ; *soum*.), et nous insisterons ici sur le traitement des données, lequel, expérimenté sur les trois groupes (lesquels présentent des variances assez satisfaisantes dans l'analyse), nous informe des proportions suivantes.

Pour le Groupe 1 :

	VBS	THEM	RHEM	CADR	OP
APOLLINE	9,00%	12,50%	5,00%	11,00%	4,00%
LEO	19,00%	16,00%	20,00%	15,00%	13,00%
MARIE SARAH	7,00%	7,00%	11,00%	19,00%	0,00%
MATHILDE	10,00%	13,00%	11,00%	9,00%	4,00%
MATTEO	25,00%	21,00%	24,00%	23,00%	8,00%
NOEMIE	30,00%	30,00%	29,00%	23,00%	71,00%

Groupe 2 :

	VBS	THEM	RHEM	CADR	OP
ALAIN	8,00%	12,00%	4,00%	0,00%	0,00%
AMBRE	31,00%	28,00%	23,00%	41,00%	34,00%
ANTONY	1,00%	2,00%	6,00%	0,00%	0,00%
KYLLIAN	16,00%	16,00%	10,00%	18,00%	12,00%
LUCAS	15,00%	12,00%	15,00%	6,00%	12,00%
MAXIM	21,00%	24,00%	25,00%	29,00%	36,00%
OPHELIO	8,00%	6,00%	17,00%	6,00%	6,00%

Groupe 3 :

	VBS	THEM	RHEM	CADR	OP
ALAIN	33,00%	44,00%	6,00%	62,00%	0,00%
JAED	9,00%	14,00%	25,00%	13,00%	0,00%
MATHIS	15,00%	14,00%	19,00%	0,00%	0,00%
SARAH	28,00%	14,00%	44,00%	25,00%	0,00%
SHON	15,00%	14,00%	6,00%	0,00%	0,00%

Ces données font ainsi apparaître qu'Alain, une fois revenu dans le troisième groupe (le sien), accapare 33 pour cent des verbes et amas verbaux du groupe, 44 pour cent des éléments thématiques, mais seulement 6 pour cent des éléments rhématiques. De même dans le groupe 2, Lucas n'emploie que 6 pour cent des cadratifs prédiqués dans le cadre des co-verbalisations, mais ses productions oscillent entre 12 et 15 pour cent des emplois recensés par ailleurs.

Confrontées au même type d'analyse que lors des précédents traitements, ces données démontrent que diverses combinaisons de variables sont possibles (la variable OP étant non représentative pour le Groupe 3, elle est considérée comme « nulle »). Or, les traitements intermédiaires que nous avons pratiqués sur ce corpus discursif ont montré que les analyses accomplies, avec là aussi des taux de variance satisfaisants, impliquent des contributions négatives aux facteurs très diverses selon les groupes, comme en témoignent les bilans ci-dessous.

Pour le groupe 1 :

	Axe 1 (+43.86%)		Axe 2 (+30.62%)	
CONTRIBUTIONS POSITIVES	RHEM	+33,0%	OP	+58,0%
	VBS	+28,0%	RHEM	+13,0%
CONTRIBUTIONS NEGATIVES	CADR	-31,0%	THEM	-9,0%
	THEM	-6,0%	VBS	-6,0%

Pour le groupe 2 :

	Axe 1 (+53.88%)		Axe 2 (+25.65%)	
CONTRIBUTIONS POSITIVES	OP	+33,0%	RHEM	+52,0%
	CADR	+24,0%	VBS	+17,0%
CONTRIBUTIONS NEGATIVES	RHEM	-6,0%	CADR	-14,0%

Pour le groupe 3 :

	Axe 1 (+60.40%)		Axe 2 (+24.86%)	
CONTRIBUTIONS POSITIVES	CADR	+35,0%	VBS	+54,0%
	THEM	+32,0%	CADR	+1,0%
CONTRIBUTIONS NEGATIVES			RHEM	-42,0%
			THEM	-1,0%

Dans le même temps, les matrices de corrélations s'avèrent particulièrement informatives, ce qui est le cas par exemple pour le groupe 1, où les contributions négatives indiquent une certaine dispersion des variables relevées :

	C1	C2	C3	C4	C5
C1 : VBS	1,000				
C2 : THEM	0,143	1,000			
C3 : RHEM	0,589	-0,327	1,000		
C4 : CADR	-0,617	0,327	-0,471	1,000	
C5 : OP	-0,151	-0,151	0,520	0,384	1,000

Variance expliquée par les composantes :

	f1	f2	f3	f4	f5
Valeur propre	2,193	1,531	1,053	0,170	0,052
% expliqué	43,866%	30,629%	21,063%	3,404%	1,038%
% cumulé	43,866%	74,495%	95,558%	98,962%	100,000%

Les représentations graphiques et les matrices de corrélations confirment le fait que de telles combinaisons de variables ne permettent pas de distinguer ce qui, dans les productions des jeunes enfants, porte spécifiquement sur les manières dont ils construisent leur discours et répondent de manière organisée aux sollicitations de l'intervenant, ou aux interventions des autres participants. Une analyse en composantes principales permettant de tester plusieurs combinaisons de variables, de manière à considérer les indices de variance les plus satisfaisants, il nous a donc paru opportun de combiner d'un côté les éléments VBS, RHEM et OP, puis de l'autre les éléments THEM, CADR et OP, avec des représentations que nous

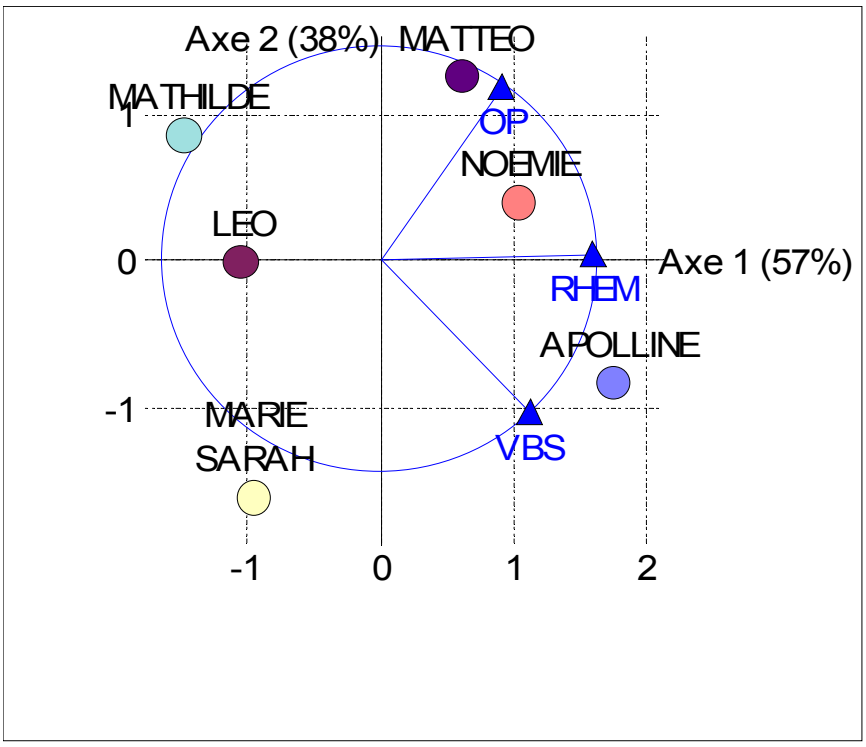
reporterons ci-après uniquement pour le premier groupe, où les productions sont pour le moins diverses.

Pour justifier ce choix en quelques lignes, disons que l'emploi des verbes et amas verbaux et celui des éléments rhématiques ont pour point commun, parmi d'autres, de ne pas porter principalement sur les manières dont les enfants encadrent, spécifient ou présentent ce à propos de quoi les informations sont données. Généralement non focaux, les VBS et RHEM constituent la plupart du temps des apports en lien avec des éléments précédemment prédiqués, ce qui n'est pas le cas de ce qui relève de la thématicité (THEM) et de l'encadrement (CADR) discursifs, lesquels coïncident le plus souvent avec des éléments qui apparaissent dans la zone préverbale. Bien entendu, la plus grande mesure s'impose dans ce domaine, et il ne s'agit pas de relever ces récurrences pour en faire autre chose que des tendances, mais ces dernières permettent toutefois de donner plus de visibilité aux démarches discursives des enfants, lesquelles démarches, à l'oral, se combinent avec de multiples autres indicateurs, au premier rang desquels interviennent les phénomènes relatifs aux données prosodiques et à la corporéité[79].

Indiquons que l'intégration de la variable des opérateurs (OP) dans les deux combinaisons de variables descriptives s'explique à partir de deux postulats simples : d'une part, les opérateurs sont des marqueurs de structuration syntagmatique, phrastique et discursive et, en tant que tels, sont proprement des polyopérateurs, en ceci qu'ils interviennent dans bien des cas sur plusieurs dimensions simultanément (Culioli 1990). De ce fait, même en les discriminant à l'appui d'indications co(n)textuelles précises, il est très difficile d'en garantir complètement le classement. D'autre part, ces opérateurs segmentaux posent des difficultés propres à leur intervention dans le cadre de corpus oraux, dans ce sens où ils s'accompagnent à l'oral de marqueurs suprasegmentaux (intonatifs : Berrendonner (2004) ou gestuels : Bouvet (2001), par exemple). Or, ces derniers se combinent avec eux et en réduisent la portée singulière.

Dans le cas du Groupe 1, les regroupements de variables apportent des informations intéressantes, dans la mesure où elles dégagent des tendances qui laissent une place à une première interprétation des données. Ci-après pour le premier regroupement (VBS, RHEM et OP) :

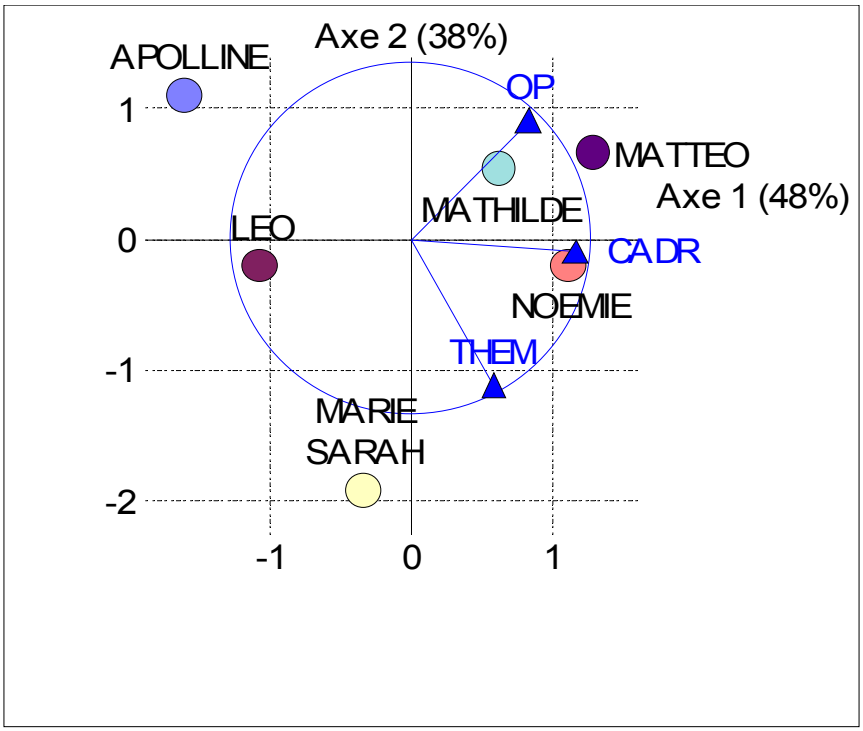
[79] *Corporéité* entendu comme « ensemble des traits concrets du corps comme être social », d'après les termes de Berthelot (1983). Ces traits sont principalement posturo-mimo-gestuels.



Contribution aux facteurs :

	Axe 1 (+57.14%)		Axe 2 (+38.31%)	
CONTRIBUTIONS POSITIVES	RHEM	+54,0%	OP	+56,0%
	VBS	+26,0%		
CONTRIBUTIONS NEGATIVES			VBS	-43,0%

Pour le deuxième regroupement (THEM, CADR et OP) :



Contribution aux facteurs :

	Axe 1 (+47.87%)		Axe 2 (+38.33%)	
CONTRIBUTIONS POSITIVES	CADR	+57,0%	THEM	+58,0%
	OP	+27,0%		
CONTRIBUTIONS NEGATIVES			OP	-40,0%

Ces représentations ainsi que les autres, par combinaisons diverses de variables, nuancent en partie la trivialité du premier traitement en données brutes, mais confirment dans le même temps ses insuffisances. Elles révèlent par ailleurs que les faiblesses, ou au contraire l'abondance des productions dans les domaines de la catégorie verbale et des éléments rhématiques masquent en partie celles qui relèvent proprement de la structuration discursive.

Dans le premier regroupement, on note que le facteur 1 rassemble les contributions des éléments rhématiques et dans une moindre mesure des éléments verbaux, et que le facteur 2 privilégie les opérateurs, au détriment notamment des éléments verbaux. Dans le deuxième regroupement, le facteur 1 rend principalement compte des cadratifs et des opérateurs, alors que le facteur 2 privilégie les éléments thématiques.

Les positionnements des variables nominatives ne peuvent se comprendre qu'à partir de ces éléments contributifs, lequels sont donc diversement explicatifs sur le plan statistique. Ils emportent néanmoins avec eux ce que nous appellerons volontiers des *suggestions*, comme dans le cas de Marie-Sarah, laquelle, en dépit de l'emploi significatif qu'elle fait des cadratifs, mais aussi du fait qu'elle verbalise autant que Mathilde et plus encore qu'Apolline, voit ses productions en partie à l'écart de celles du noyau du groupe, à la fois en termes de production et de structuration discursives.

Cela n'est vraisemblablement pas le cas, en l'occurence, d'Apolline et de Mathilde, qui en termes de proportions dans le groupe, se positionnent différemment selon les variables expliquées. Or, nous retrouvons ici l'absence notamment d'une « métrique simple » des composantes principales, lesquelles ne permettent pas, *a priori*, de formuler des déductions robustes à partir des représentations graphiques basées sur des instantanés.

5 QUELQUES CONCLUSIONS

Les conclusions que nous tirerons des applications présentées *supra* sont donc surtout d'ordre méthodologique. La première est que, dans le cadre des analyses multivariées de ce type de corpus discursif sous forme d'ACP, les positionnements des variables nominatives, dans les représentations graphiques, renvoient plus à des profils de production qu'à des profils d'acquisition, même si l'ACP permet de rendre plus concrets les rapprochements et les écarts entre les variables descriptives, auxquels elle apporte des garanties numériques. La deuxième est que l'ensemble de ces éléments confortent en partie le point de vue suivant lequel il convient de distinguer ce qui concerne les productions en données brutes par rapport à ce qui concerne spécifiquement la structuration discursive. En la matière, le choix qui consiste à regrouper d'un côté les verbes et amas verbaux avec les éléments rhématiques, et de l'autre les éléments thématiques avec les cadratifs, à partir du moment où le traitement appliqué au corpus étiqueté fait l'objet d'une analyse multivariée, suppose de multiplier les tests pour déterminer quelles peuvent être les différentes contributions des opérateurs.

En marge du coût opérationnel que représentent ces types de traitements, la principale objection que l'on pourrait formuler ici est que l'analyse multivariée appliquée sur des corpus instantanés demeure tout à fait incomplète si elle n'est pas confortée par une version longitudinale. En effet, les déductions qu'il est possible d'envisager à partir d'instantanés, non seulement se révèlent intermédiaires, mais aussi nécessitent des recoupements constants dans les informations statistiques et leurs corrélats informatiques, avec tout ce que les représentations graphiques correspondantes comportent d'apriorisme.

Les corpus longitudinaux, en relatant les productions verbales de groupes restreints d'enfants collectées sur plusieurs mois ou plusieurs années, en plus du fait qu'ils apportent davantage d'indications dans le domaine de la composition du lexique (Bassano, Eme et Champaud 2005), contribuent également à mesurer les acquisitions des enfants à travers la « grammaire » qu'ils mettent en oeuvre et l'organisation discursive qui s'établit dans leurs interventions. D'autre part, de par leur volume et leur diversification, ils font l'objet d'un suivi régulier et permettent de valoriser des tendances et des constantes qui ne peuvent être que pressenties à partir de corpus instantanés.

Ces derniers nous semblent toutefois productifs à quelques égards. En effet, les instantanés permettent de pratiquer une analyse exploratoire du corpus, et d'effectuer une première approche de données textuelles pour une part converties en variables numériques. En outre, ils conduisent à mesurer en partie la représentativité du corpus et à en aborder diversement la matérialité. Ce type d'apport s'avère par conséquent plus opportun au moment du tri parmi les démarches analytiques possibles, qu'à celui du traitement informatique des données. A ce titre, ils représentent quand même un gain de temps non négligeable, d'autant qu'ils contribuent à dégager ce qui, dans les démarches envisagées, concerne plus particulièrement les difficultés de mise en œuvre.

6 RÉFÉRENCES

Apothéloz D. (1997). « Les Dislocations à gauche et à droite dans la construction des schématisations ». Dans A. Berrendonner et D. Miéville (éds.), *Logique, Discours et Pensée. Mélanges offerts à Jean-Blaise Grize*. Bern : Lang. p. 183-217.

Balthasar L., Bruxelles S., Mondada L et Traverso V. (2003). « *Attends ça fait travailler le cerveau :* usages et tendances à la grammaticalisation de *attends* en français parlé en interaction ». Dans *Linguistique de Corpus, 36e colloque de la Societas Linguistica Europea.* Lyon. p. 4-7.

Bassano D., Eme E. et Champaud C. (2005). « A naturalistic study of early lexical development : General processes and inter-individual variations in French children ». *First Language*, 25 (1), p. 67-101.

Benzitoun C. (2004). « L'Annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? ». Dans *Actes du Colloque TALN 2004*, URL : http://aune.lpl.univ-aix.fr/jep-taln04/proceed/actes/recital2004/Benzitoun.rec04.pdf [consulté le 25 avril 2007].

Berrendonner A. (2004). « Grammaire de l'écrit *vs* grammaire de l'oral : le jeu des composantes micro- et macro-syntaxiques ». Dans A. Rabatel (éd.), *Interactions orales en contexte didactique*. Lyon : PUL. p. 249-264.

Berthelot J. M. (1983). « Corps et Société. Problèmes méthodologiques posés par une approche sociologique du corps ». *Cahiers internationaux de sociologie*, LXXXIV, p. 119-131.

Bouvet D. (2001). *La Dimension corporelle de la parole*. Louvain : Peeters.

Charolles M., Le Draoulec A., Pery-Woodley M. P. et Sarda L. (2005). « Temporal and spatial dimensions of discourse organisation ». *Journal of French Language Studies,* 15-2, p. 203-218.

Cibois P. (2007). *Les Méthodes d'analyse d'enquête*. Paris : PUF.

Cuenca M. J. et Hilferty J. J. (1999). *Introducción a la linguística cognitiva*. Barcelona : Editorial Ariel.

Culioli A. (1990). *Pour une linguistique de l'énonciation*, 1. Paris : Ophrys.

Desclés J. P. (2008). « Opérations de prédication et de détermination ». *Lidil*, 37, p. 61-98.

Dister A. et Simon A.C. (2008). « La Transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé ». *Arena Romanistica* 1/1, p. 54-79.

Gerdes K. et Kahane S. (2006). « L'Amas verbal au cœur d'une modélisation topologique de l'ordre des mots ». *Linguisticae Investigationes*, 29-1, p. 75-89.

Grobet A. (1997). « La Ponctuation prosodique dans les dimensions périodique et informationnelle du discours ». *Cahiers de Linguistique française* 19, p. 83-123.

Mac Whinney B. (2000). *The CHILDES project : Tools for analysing talk* (troisième édition). Mahwah NJ : Lawrence Erlbaum.

Martin O. (1997). « Aux origines des idées factorielles. Des théories aux méthodes statistiques ». *Histoire & Mesure*, 12 (3-4).

Mel'čuk I. (2001). *Communivative Organization in Natural Language*. Amsterdam/Philadelphia : Benjamins.

Pincemin B. (2007). « Introduction » au numéro 6 de la revue *Corpus* : 5-15 [article consulté le 6 juillet 2009] http://corpus.revues.org/index812.html

Rastier F. (2009). « Quantité et Qualité en sémantique de corpus ». *Comm. aux 6èmes Journées de Linguistique de corpus*. Lorient : Université de Bretagne-Sud.

Rebuschi G. (2002). « Coordination et subordination : vers la co-jonction généralisée ». *Bulletin de la Société de Linguistique de Paris,* 97-1, p. 37-94.

Sansonetti L. (2003). « Approche lexicométrique de corpus d'interactions verbales entre un adulte et un enfant en cours d'acquisition du langage. Résultat d'expérience ». Dans G. Williams (éd.), *Actes des Troisièmes Journées de la Linguistique de Corpus*. Lorient : Université de Bretagne-Sud [consulté le 30 mai 2009] http://web.univ-ubs.fr/corpus/jlc3/1_4_sansonetti.pdf

Saussure L. de et Sthioul B. (2002). « Interprétations cumulative et distributive du connecteur *et* : temps, argumentation, séquencement ». *Cahiers de linguistique française,* 24, p. 293-314.

Shriberg E., Stolcke A., Hakkani-Tur D. et Tur G. (2000). « Prosody-Based Automatic Segmentation of Speech into Sentences and Topics ». *Speech Communication* 32-1, p. 127-154.

Sinclair J. (2005). « Meaning in the Framework of Corpus Linguistics ». Dans W. Teubert (éd.), *Lexicographica.* Tubingen : Niemeyer, p. 20-32.
Torterat F. (2002). *Approche des invariants de quelques joncteurs en français : pour une complémentarité des termes de* coordination *et de* jonction. Th¯se de doctorat, Paris-Sorbonne.
Torterat F. (2010). « Le Recours aux corpus discursifs : difficultés et possibilités pratiques ». Dans M. St. Rădulescu (éd.), *La Méthodologie pour un apprentissage de la recherche*, Chisinau : Université de Ion Creangă.
Torterat F. (*soum*.). « Analyse de quelques variabilités interindividuelles dans le domaine de la structuration discursive : à partir d'un corpus caractérisé ». Dans *Actes du Colloque « Linguistic Approaches to Text Structuring »*. Paris : ENS.
Wilkins D. P. (1992). « Interjections as deictics ». *Journal of Pragmatics* 18 / 2-3, p. 119-158.

AUTOUR DU PROJET SCIENTEXT : ÉTUDE DES MARQUES LINGUISTIQUES DU POSITIONNEMENT DE L'AUTEUR DANS LES ÉCRITS SCIENTIFIQUES

Agnès Tutin[*], Francis Grossmann[*], Achille Falaise[+], Olivier Kraif[*]
[*]Lidilem – Université Grenoble 3 –Stendhal
agnes.tutin@u-grenoble3.fr
francis.grossmann@u-grenoble3.fr
olivier.kraif@u-grenoble3.fr
[+]GETALP – LIG – Université Grenoble 1 – Joseph Fourier
achille.falaise@imag.fr

Nous dédions cet article à la mémoire de notre jeune collègue Robert Barr, disparu brutalement et prématurément en avril 2010.

RÉSUMÉ

Cet article présente le projet Scientext, qui a permis de constituer un corpus d'écrits scientifiques variés et des outils logiciels permettant d'effectuer une étude linguistique du positionnement et du raisonnement dans les écrits scientifiques, à travers des marques linguistiques. Nous retraçons ici les ressources développées pour le français dans le cadre du projet et présentons une étude de cas du positionnement de l'auteur, à travers l'étude des verbes de positionnement associés à un sujet auteur en sciences humaines.

1 INTRODUCTION

Le projet ANR Scientext[1] s'inscrit dans le domaine de la linguistique de corpus. Il poursuit deux objectifs : d'une part, constituer un corpus d'écrits scientifiques variés, permettant à la fois de comparer des ensembles de disciplines et des sous-genres scientifiques, comme la thèse ou l'article de recherche ; d'autre part, l'étude linguistique des marques explicites du positionnement et du raisonnement de l'auteur dans ce genre.

Nous décrivons ici dans un premier temps les objectifs du projet, les corpus constitués à cette fin et les annotations effectuées. Dans un second temps, nous abordons une problématique linguistique traitée dans le cadre du projet, l'étude des verbes de positionnement explicite en sciences humaines (linguistique, psychologie, sciences de l'éducation). Notre objectif est ici d'observer dans quelle mesure la variable disciplinaire apparaît déterminante dans l'emploi de la première personne et de ces marques de positionnement explicites, et parallèlement d'observer comment se construit la visibilité de l'auteur scientifique. Nous présentons enfin les modes d'exploitation du corpus mis en œuvre dans le site en ligne où, à la façon de Frantext, les corpus sont librement interrogeables par divers types de requêtes (« simples », « sémantiques », « avancées").

[1] Projet dans le cadre de l'ANR « Corpus et outils de la recherche en sciences humaines et sociales » (2007-2010). Site du projet : http://scientext.msh-alpes.fr.

2 PRÉSENTATION DU PROJET SCIENTEXT

2.1 Objectifs du projet

Le projet Scientext vise à produire des données et des analyses linguistiques sur les écrits scientifiques intéressant les linguistes, mais aussi les spécialistes de l'extraction d'information qui cherchent à identifier des passages spécifiques, comme les références à autrui (par exemple, Siddhartan & Teufel, 2007), véritable enjeu dans la veille technologique. A plus long terme, le projet pourrait également permettre la construction de bases de données lexicales, d'outils d'aide à la rédaction basés sur corpus (Kraif & Tutin 2009), ainsi que l'élaboration d'outils pour l'extraction d'informations scientifiques.

Scientext s'inscrit pleinement dans la linguistique de corpus, puisqu'il cherche à mettre en évidence les spécificités lexicologiques et énonciatives du genre des écrits scientifiques en se basant sur un ensemble de textes authentiques. Le projet recourt en outre aux techniques du traitement automatique des langues, à la fois pour l'analyse syntaxique du corpus (effectué à l'aide de l'analyseur de dépendance Syntex développé par Didier Bourigault (2007)), et pour l'interrogation du corpus qui exploite des requêtes complexes (Cf. Kraif 2008 ; Falaise et Tutin 2010), comme on le verra dans la troisième section. L'analyse du discours, la lexicologie et l'approche énonciative sont aussi convoquées pour l'analyse linguistique du positionnement et du raisonnement.

Trois équipes ont été impliquées dans ce projet. Le LIDILEM[2] (Laboratoire de Linguistique et de Didactique du Française Langue Etrangère et Maternelle, Université Grenoble 3-Stendhal) qui a coordonné le projet, a constitué le corpus d'écrits scientifiques français et l'interface informatique permettant d'exploiter le corpus. Le laboratoire Littérature Langage Société[3], de l'Université de Chambéry, a élaboré un corpus d'anglais académique de locuteurs non natifs. L'équipe LiCorn (Linguistique de Corpus, Université de Bretagne Sud[4]), a recueilli et traité un corpus d'anglais scientifique, principalement en sciences de la vie et médecine.

Le produit final réalisé est un site Web (adresse : http://scientext.msh-alpes.fr), à la façon de Frantext, permettant de sélectionner un corpus d'après une combinaison de critères (disciplines, sous-genre textuel, parties textuelles). Le corpus est interrogeable à l'aide de requêtes linguistiques simples ou complexes, à partir des étiquettes morphosyntaxiques ou de relations syntaxiques de dépendance. Des grammaires locales ont été établies sur les thèmes du positionnement et du raisonnement et permettent d'accéder au texte par le biais de recherches sémantiques spécifiques, par exemple sur les verbes d'opinions ou les formulations d'hypothèse.

La problématique traitée porte sur l'expression linguistique du positionnement et du raisonnement de l'auteur. A travers le positionnement, l'auteur s'inscrit comme sujet par rapport à ses devanciers, à ses contemporains, il définit sa spécificité, ses choix, comme nous le verrons plus en détail dans la deuxième section. L'étude du raisonnement permet de retracer son cheminement intellectuel, ce sur quoi il s'appuie et les déductions qu'il opère. Nous souhaitons en outre étudier les différents types de variation que l'on observe en fonction du genre textuel et de la discipline, des disparités importantes ayant été observées pour ce second paramètre. Fløttum *et al*. (2006), en étudiant les articles de recherche en médecine, linguistique et économie dans trois langues (anglais, français et norvégien) ont ainsi mis en évidence que le paramètre

[2] Personnes impliquées au LIDILEM : F. Grossmann, A. Tutin (responsables), G. Antoniadis, F. Boch, C. Cavalla, M. Florez, O. Kraif, I. Novakova, M. Mroué, M.L. Nguyen, F. Rinck.

[3] Personnes impliquées au LLS : J. Osborne, A. Henderson, R. Barr.

[4] Participants : Geoffrey Williams, Chrystel Millon.

disciplinaire était plus déterminant que la langue ou la culture du chercheur. A l'aide de notre corpus et des outils développés, nous souhaitons étendre ces observations à d'autres disciplines, d'autres genres textuels et d'autres paramètres linguistiques.

2.2 Constitution du corpus français d'écrits scientifiques

Le projet Scientext intègre trois grands corpus :

- Un corpus d'écrits scientifiques du français, pluridisciplinaire, et représentant des genres variés, qui contient un peu moins de 5 millions de mots.
- Un corpus anglais d'apprenants, comprenant des travaux longs d'étudiants en anglais langue étrangère (1,1 million de mots).
- Un corpus anglais d'écrits scientifiques, tiré du corpus BMC, principalement en biologie et en médecine, qui avoisine 13 millions de mots, qui a fait l'objet d'études lexicologiques (Williams & Millon, à paraître).

Seul le corpus français sera ici décrit en détail, mais les autres corpus sont annotés selon les mêmes principes (Cf. Henderson *et al.* 2009). Pour étudier les points linguistiques que nous souhaitions explorer, les marques du raisonnement et du positionnement de l'auteur, nous avons constitué pour le français un corpus d'écrits scientifiques diversifié, aussi bien en ce qui concerne le sous-genre (articles scientifiques, communications écrites, thèses ou mémoires d'habilitation à diriger des recherches) que les disciplines. Il était bien entendu exclu d'inclure dans le présent projet la totalité ou la quasi-totalité des disciplines représentées, par exemple par les différentes sections du CNRS[5] ou des sections du CNU[6]. Nous avons ainsi sélectionné des disciplines qui nous paraissaient représentatives de familles scientifiques plus larges et pour lesquelles les écrits étaient facilement disponibles. Trois familles de disciplines sont incluses : les sciences humaines (linguistique, psychologie, sciences de l'éducation et dans une certaine mesure, le traitement automatique des langues), les sciences expérimentales (biologie, médecine) et les sciences appliquées ou sciences pour l'ingénieur (électronique, mécanique), les frontières entre ces familles n'étant pas étanches. Les sous-genres sélectionnés intègrent des articles de recherche, des communications écrites[7], des thèses de doctorat et des mémoires d'habilitation à diriger les recherches[8]. Le corpus public, consultable en ligne, compte à peu près 5 millions de mots[9]. Le tableau 1 présente le détail du corpus dont on peut relever immédiatement qu'il n'est pas véritablement équilibré : les sciences humaines y sont surreprésentées, en particulier pour les articles, genre absent pour les sciences appliquées où la langue dominante est l'anglais. Les articles obtenus en médecine et biologie sont extraits d'une revue de très bonne qualité *Médecine/Science* qui, sans être une revue de vulgarisation, a néanmoins pour objectif de diffuser au plus grand nombre les recherches récentes dans ce domaine dans la francophonie.

---

[5] Liste des sections du Centre National de la Recherche Scientifique : http://www.cnrs.fr/comitenational/sections/intitsec.htm.

[6] Groupes et sections du CNU : http://www.cpcnu.fr/sectionsCnu.htm.

[7] La liste complète des revues et des conférences est donnée en annexe.

[8] La liste complète des textes ne peut pas être indiquée ici, mais elle est consultable en ligne sur le site de Scientext.

[9] Une autre partie du corpus, pour laquelle les droits n'ont pas été obtenus, est consultable sur notre Intranet, avec un mot de passe.

	Articles et communications écrites	Thèses de doctorat	Mémoires d'Habilitation à Diriger des Recherches
Linguistique	67 textes	8 textes	4 textes
Psychologie	12 textes	5 textes	1 texte
Sciences de l'Education	55 textes	6 textes	1 texte
Traitement Automatique des Langues	13 textes	4 textes	
Total Sciences Humaines	147 textes	23 textes	6 textes
Biologie	10 textes	11 textes	
Médecine	12 textes	2 textes	
Total Sciences expérimentales	22 textes	13 textes	
Electronique		5 textes	
Mécanique		2 textes	1 texte
Total Sciences Expérimentales		7 textes	1 texte

Tableau 1 : La composition du corpus français public dans Scientext

Les corpus ont été annotés au plan structurel en suivant les recommandations de la Text Encoding Initiative (TEI Lite, P5[10]), en isolant les différentes parties textuelles de l'article : résumé, introduction, corps du texte, conclusion, remerciements, notes de bas de page, bibliographie, annexes, titres. Ce travail de balisage a été automatisé lorsque cela apparaissait possible, mais la tâche a souvent dû être complétée manuellement, ce qui s'est révélé extrêmement fastidieux, et a nécessité le recrutement de nombreux vacataires qui ont utilisé à cette fin des outils spécialisés[11]. L'annotation des parties textuelles est très utile pour mener des études linguistiques fines sur les résumés, les introductions ou les notes de bas de page qui présentent des spécificités manifestes en ce qui concerne le positionnement et le raisonnement. La mise en forme (gras, italique, structure de liste) a été conservée lorsqu'elle pouvait être générée automatiquement, mais pas de façon systématique car le balisage a souvent été réalisé à partir d'un format texte où ces informations avaient été effacées[12]. En outre, le corpus a été analysé linguistiquement (et automatiquement) à l'aide de l'analyseur syntaxique de dépendance Syntex, développé par Didier Bourigault (2007)[13], dont les performances ont été soulignées dans la campagne d'évaluation EASY[14]. Pour chaque phrase, ont été indiqués pour chaque mot le lemme, la catégorie syntaxique et les relations de dépendance qui le lient aux autres mots de la phrase. La figure 1 montre ainsi un exemple pour l'analyse de la phrase : *Enfin, nous avons fait*

[10] http://www.tei-c.org/Guidelines/P5/.

[11] En particulier, le logiciel Oxygen : http://www.oxygenxml.com.

[12] Une partie du corpus annotée au plan structurel est disponible pour des fins de recherche. Une convention « creative commons » a été signée à cette fin avec les auteurs et les éditeurs qui ont accepté de rendre leurs corpus disponible. Pour obtenir le corpus, il faut contacter les responsables du projet (scientext@u-grenoble3.fr) et signer une convention.

[13] Que nous remercions très chaleureusement ici.

[14] Voir les détails sur : http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=bourigault&subURL=syntex.html

l'hypothèse que les élèves n'occupaient pas une place aléatoire au sein des 4 classes et qu'elle était en relation avec leur statut scolaire.

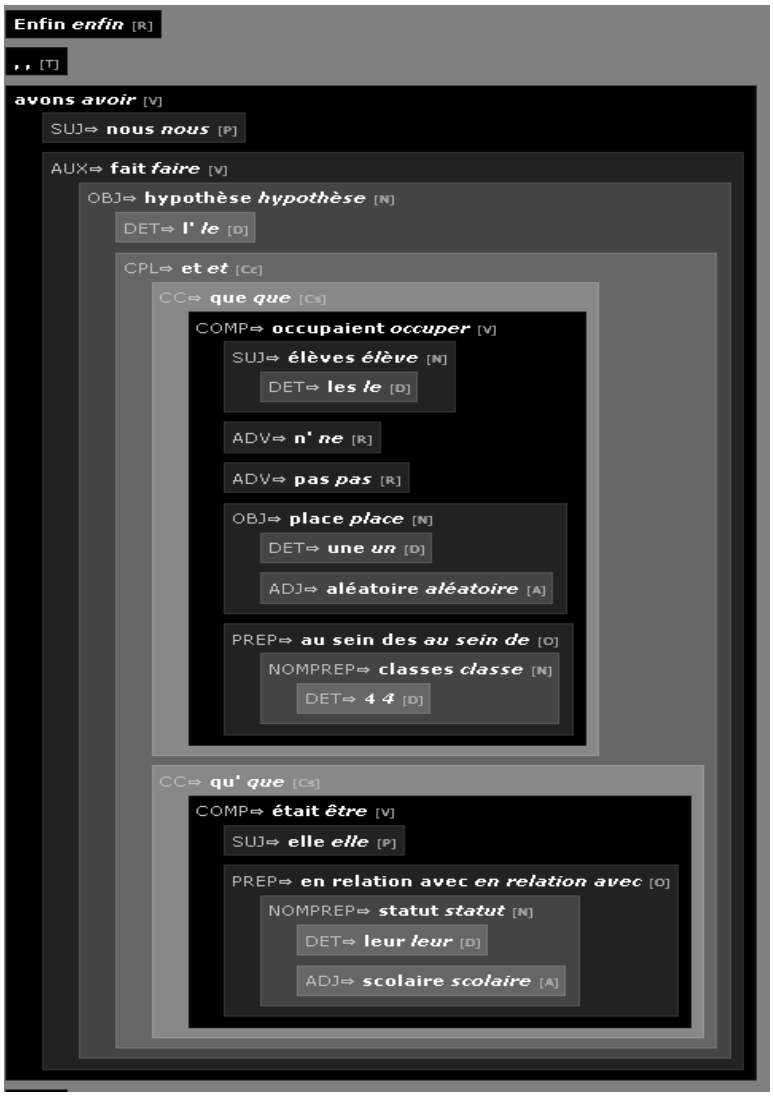


Figure 1 : Analyse syntaxique à l'aide de Syntex (Bourigault 2007) de « Enfin, nous avons fait l'hypothèse que les élèves n'occupaient pas une place aléatoire au sein des 4 classes et qu'elle était en relation avec leur statut scolaire. »

On repère dans cette analyse, à côté des mots sous leur forme fléchie, les lemmes, les catégories syntaxiques et une analyse de dépendance de surface entre les éléments. Par exemple, l'auxiliaire *a* est relié à *fait* par une relation *aux*. Pour une approche plus sémantique (Cf section 3.2.1), il faudra donc recalculer la relation syntaxique « profonde » entre *on* et *fait* dans *on a fait* à partir des relations de surface. L'utilisation de cet analyseur syntaxique permet néanmoins de créer des requêtes et des grammaires générant peu de bruit et peu de silence, en tout cas bien plus performantes qu'un simple étiquetage morpho-syntaxique.

3 LE POSITIONNEMENT DANS L'ÉCRIT SCIENTIFIQUE

3.1 Comment définir le positionnement ?

Un des deux thèmes linguistiques développés dans notre projet est celui du positionnement de

l'auteur dans les écrits scientifiques[15]. Le positionnement n'est pas un concept linguistique, contrairement à la notion de « point de vue », utilisée dans les linguistiques de l'énonciation dans le cadre de l'étude de la polyphonie énonciative (par exemple, Nølke *et al.* 2004 ; Rabatel 1988). Dans le cadre de Scientext, nous définissons le positionnement comme la façon dont l'auteur s'inscrit dans une communauté de discours, comment il s'évalue et évalue ses pairs, et quelles propositions propres il met en avant. Ce thème permet d'aborder la figure de l'auteur dans ce type d'écrits, ainsi, plus largement que celle de l'*auctorialité scientifique* à travers l'étude de marques énonciatives, lexicales et syntaxiques spécifiques.

L'étude des marques linguistiques du positionnement dans les écrits scientifiques permet donc d'embrasser trois aspects différents, bien que complémentaires :

- La question du *contexte scientifique, du cadre théorique et des références* propres à un auteur ou à une équipe. Il peut s'agir de la filiation intellectuelle, c'est-à-dire l'approche, les idées, voire la terminologie dont un auteur s'inspire, ou le cadre théorique dans lequel il s'inscrit explicitement (Boch. F *et al*. 2009; Garcia P.P. 2008; Grossmann *et al.* 2009). Cette problématique inclut également l'ensemble des références à autrui « neutres » qui apparaissent dans l'écrit scientifique, même si le positionnement de l'auteur par rapport aux auteurs mentionnés apparaît ici moins explicite[16].

- Dans le sens plus restreint de « prise de position », l'étude du positionnement permet d'observer *les moyens linguistiques utilisés pour exprimer un parti pris, un jugement ou une évaluation*. Cela peut concerner l'évaluation d'un point théorique, d'un résultat, d'une démarche (Cf. Tutin 2010a, Cavalla & Tutin, à paraître), sur la démarcation ou la distance vis à vis des pairs (*contrairement à X ... nous nous différencions de X*) (Chavez 2008), ou bien au contraire l'adhésion ou la convergence de vue. L'évaluation peut aussi concerner la démarche de l'auteur, dans la formulation de la conformité/non-conformité aux attentes (Rinck *et al.* 2007) ;

- Enfin, le positionnement concerne les *choix propres, les propositions et les déductions opérés* par l'auteur, par le biais de l'emploi de la première personne (*nous optons pour ... nous concluons que ... nous avons montré que*...) ou l'emploi d'un lexique déontique (*il faut ... ce problème doit à nouveau être traité* ...).

Dans ce projet, nous avons fait l'hypothèse que l'expression du positionnement était relativement stéréotypée et s'exprimait à travers une phraséologie repérable, qui pouvait être traitée dans des grammaires locales (de l'opinion, de l'évaluation, de la démarcation …). Nous avons aussi supposé que ces expressions récurrentes appartenaient à un sous-langage plus spécifique du genre que des disciplines en tant que telles, un lexique scientifique transdisciplinaire (Tutin 2007) de la langue scientifique générale au sens de Pecman (2004).

---

[15] La question du raisonnement ne sera pas abordée dans cet article.

[16] Dans le cadre du projet Scientext, nous distinguons la citation ou la référence à autrui « neutre » qui n'indique pas de positionnement explicite de l'auteur par rapport à la référence citée, de la citation ou référence à autrui « positionnée » qui indique une position explicite de l'auteur (Ex : *contrairement à Duschmoll (1970) ... Nous reprenons le modèle de Duschmoll (1970)*) ... La citation neutre n'est généralement pas intégrée syntaxiquement comme en (1), alors que la citation positionnée est intégrée syntaxiquement comme en (2) (Cf. Florez, 2010) :

(1) Les travaux sur le positionnement dans les écrits scientifiques sont souvent d'inspiration énonciative (Cf. Fløttum *et al.* 2006).

(2) Contrairement à Fløttum (2006), notre approche du positionnement n'est pas exclusivement énonciative.

Bien entendu, toutes les citations intégrées ne sont pas positionnées.

3.2 L'exemple du lexique verbal du positionnement

Plusieurs études ont été réalisées sur le thème du positionnement dans le cadre du projet Scientext. Nous reprenons ici un exemple représentatif des études sur ce thème, l'utilisation du lexique verbal marquant explicitement un engagement de l'auteur (Tutin, 2010b). Les écrits scientifiques sont souvent considérés comme un genre « neutre », avec un fort effacement énonciatif, où l'auteur se dissimulerait derrière la présentation de faits objectifs et des modalités de raisonnement partagés par la communauté scientifique. Les travaux accomplis sur ce sujet dans les dernières années (par exemple, Swales 1990 ; Hyland 2002 ; Fløttum *et al*. 2006 ; Rinck 2006) montrent cependant qu'il n'en est rien, en tout cas dans certaines disciplines, et que l'écrit scientifique est véritablement un texte argumentatif où la dimension rhétorique est fortement présente. Cette étude, qui comparait trois disciplines des sciences humaines et sociales, la linguistique, la psychologie et les sciences de l'éducation[17], cherchait à mettre en évidence les modalités d'engagement explicites de l'auteur à travers les verbes de positionnement associés à un pronom sujet (Ex : *je cherche à démontrer, nous pensons que..*). Nous faisions l'hypothèse que la présence auctoriale s'établit diversement selon les disciplines des sciences humaines. On peut imaginer, comme mis en évidence par Fløttum *et* al. (2006), qu'elle sera assez manifeste en sciences du langage où l'auteur cherche souvent à développer une pensée ou un modèle propre. En psychologie cognitive et sociale, en revanche, on peut supposer que les écrits, qui se rapprochent par la structure IMRaD[18] et les méthodes (expérimentales) des sciences dures, pourraient ainsi en adopter le style plus « objectif », avec moins de références explicites aux auteurs de l'article et l'emploi moins marqué de verbes exprimant un point de vue explicite. En outre, on peut aussi s'attendre à ce que le type de verbe utilisé soit fortement lié à la valeur référentielle du pronom sujet, selon qu'il renvoie strictement à l'auteur ou qu'il intègre aussi la communauté de discours.

L'étude a été réalisée partir d'un corpus de 60 articles de linguistique, psychologie et sciences de l'éducation[19] (3x20), en observant de façon systématique dans les introductions et les conclusions les verbes qui engagent fortement l'auteur, associés à un pronom auteur repéré semi-automatiquement (*je, nous, on*).

Ont été retenus comme verbes de positionnement explicites :

- des verbes qui expriment une **opinion** ou un **point de vue** (*penser, croire, considérer que, juger*…), ou une distance/adhésion par rapport aux pairs (*se distinguer de, rejoindre*…), ou à un questionnement (*se demander*…)
- des verbes indiquant un **choix** (*choisir, retenir, opter pour*…), une **intention** (*vouloir, souhaiter, projeter*…) ou des hypothèses (*faire, formuler, émettre une hypothèse ; supposer*).
- des verbes qui indiquent un apport **apport spécifique de l'auteur**, une **proposition** (*proposer* …), une **preuve** ou une **démonstration** (*montrer, prouver* ...) ou bien des **résultats** (*dégager, souligner* …).

Les résultats de cette étude ont permis de mettre en évidence deux grandes tendances. La première traduit avant tout une visibilité assez modérée de l'auteur. S'il ne se cache pas (il apparaît à la première personne, voire même à la première personne du singulier en linguistique),

[17] D'autres disciplines sont abordées dans le cadre du projet Scientext, mais nous avons choisi de nous limiter ici à un sous-ensemble de sciences humaines et sociales, Fløttum *et al*. (2006) ayant déjà montré que la présence auctoriale était faible dans les sciences expérimentales comme la médecine.

[18] IMRaD : Introduction, Méthodes, Résultats, Analyse, Discussion. C'est un plan textuel imposé dans les disciplines expérimentales.

[19] Certains des articles analysés ici n'appartiennent pas au corpus public présenté en 1.

dans les textes examinés, l'auteur se manifeste discrètement. Ainsi, le *nous* de modestie est partout préféré. Bien que ce *nous* conventionnel soit souvent difficile à interpréter de façon univoque, l'auteur semble mettre par cet emploi l'accent sur son appartenance à une communauté de discours : il met peu en avant son individualité et sa spécificité. En outre, les verbes de positionnement employés ne sont pas majoritairement des verbes à « fort » positionnement (comme les verbes d'opinion), comme on l'observe sur la figure 2, mais plutôt des verbes qui indiquent les choix effectués par l'auteur (Ex : *nous avons opté ...*) ou les apports scientifiques de la recherche effectuée (Ex : *Nous avons dégagé ...*). Enfin, on relève que les verbes à « fort positionnement, comme *penser*, tendent à être modalisés dans des formules du type *on peut penser que ...* On pourrait voir dans ce type de modalisation une prise de risque minimale de l'auteur, qui ne cherche pas ici à s'engager dans une opinion affirmée, mais on peut surtout interpréter ce type de formulation, à l'instar de Hyland (1998), comme une forme de négociation avec le lecteur (le *on* inclut ici le lecteur). L'auteur ne cherche pas ici à imposer son point de vue (ce n'est pas le mode de fonctionnement de la « Science ») mais il montre qu'au vu des résultats obtenus, tout chercheur (y compris l'auteur et le lecteur), tirerait des conclusions identiques. En bref, ces éléments semblent montrer que l'auteur s'inscrit avant tout fortement dans la communauté de discours, et met peu en avant son individualité.

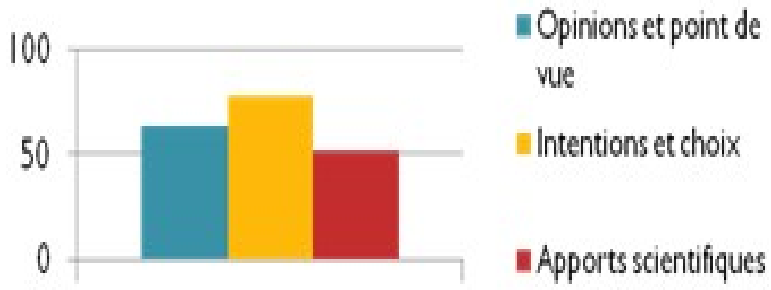


Figure 2 : Répartition des différents types de verbes de positionnement

La deuxième grande tendance observée est, comme observé par Fløttum *et al.* (2006) dans d'autres familles de disciplines, une forte variabilité disciplinaire au sein de ces trois disciplines des sciences humaines, en tout cas dans ce corpus.

Comme on l'observe dans la figure 3, la proportion des verbes de positionnement par rapport à l'ensemble des verbes va ainsi de 1 pour la psychologie à 3 pour la linguistique.

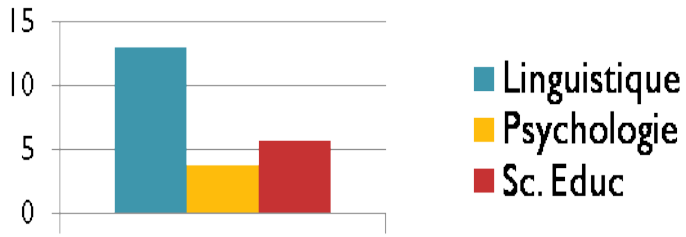


Figure 3 : Répartition disciplinaire des verbes de positionnement

Si l'on entre davantage dans les détails, on s'aperçoit que la linguistique se caractérise par une forte proportion de verbes de positionnement, de toutes sortes, mais en particulier d'opinions, intentions, résultats et démonstrations, ainsi que par la présence notable du *je* (que l'on rencontre peu ailleurs). En sciences de l'éducation, la présence des verbes de positionnement est plus modérée, et l'accent est mis sur la justification de la démarche, à travers les opinions et les intentions. En psychologie enfin, on dénombre encore moins de verbes de positionnement et ceux-ci apparaissent surtout pour indiquer une opinion et une démarche expérimentale.

On voit ici que l'étude linguistique menée, qui doit bien sûr être complétée, permet de mettre en évidence certains éléments essentiels du fonctionnement des disciplines, comme les critères de scientificité et d'évaluation propres à chacune. Ainsi, la linguistique semble mettre l'accent sur une forme d'individualité et de créativité de l'auteur, alors qu'en sciences de l'éducation, les raisons d'être (sociales ?) de la recherche sont soulignées. La psychologie, qui se rapproche sur ce plan des sciences expérimentales, met plutôt l'accent sur les hypothèses et les résultats obtenus. Il faut cependant se garder de conclusions trop hâtives : les études linguistiques ont montré que la question du positionnement était plurielle et que toutes ses dimensions ne convergeaient pas nécessairement. Ainsi, si la linguistique se caractérise par une certaine visibilité de l'auteur, les écrits de cette discipline présentent par ailleurs peu de marqueurs de positionnement par rapport aux pairs. En économie, en revanche, on observe la tendance inverse avec des stratégies de persuasion plus offensives en direction de pairs à travers les marques d'évaluation (Tutin 2010) ou de filiation (Grossmann *et al.* 2009) mais une visibilité de l'auteur moindre. Des études de cas sur le corpus Scientext sur plusieurs points linguistiques (citation positionnée, propositions propres de l'auteur ...) restent à accomplir pour brosser un portait nuancé et adéquat du positionnement.

4 PRÉSENTATION DU SITE SCIENTEXT : MODES D'EXPLOITATION DES CORPUS

Dans le cadre de notre projet, outre la constitution de corpus, un outil d'exploitation des corpus annotés a été élaboré par Achille Falaise sous la forme d'un site Internet (adresse : http://scientext.msh-alpes.fr), librement consultable, pour interroger les corpus, entre autres, sur les marques du positionnement et du raisonnement à l'aide de grammaires prédéfinies. L'exploitation du corpus se fait en trois étapes. L'utilisateur sélectionne le corpus dans un premier temps, puis effectue sa requête (« sémantique », « libre et guidée », ou « avancée »). Enfin, il affiche les résultats sous formes de concordances ou de traitements statistiques simples.

4.1 Sélection des textes

La première étape consiste à sélectionner le corpus, à la façon de Frantext, selon un ensemble de paramètres, comme indiqué sur la copie d'écran figure 4. L'usager peut choisir la ou les discipline(s), les genres textuels et les parties textuelles, grâce au balisage structurel réalisé. Il pourra par exemple sélectionner les résumés des articles et communications des sciences humaines.

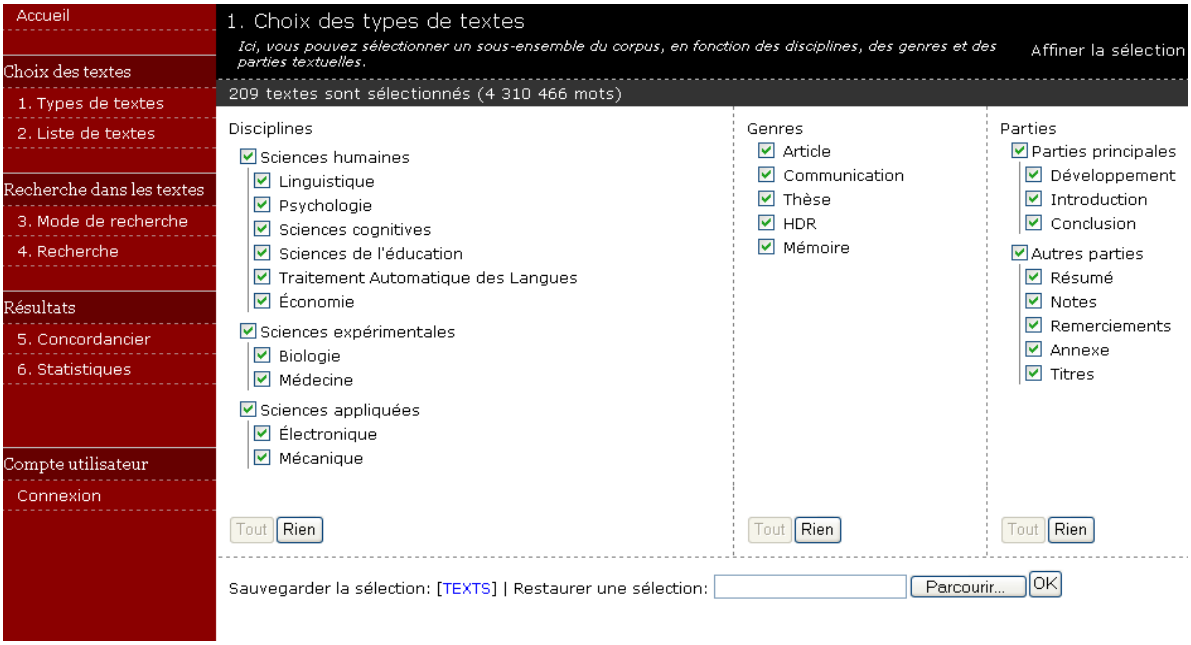


Figure 4 : Sélection du corpus dans Scientext

Une fois les textes correspondant aux choix affichés, l'usager peut ensuite affiner la sélection en sélectionnant ou non les textes un par un. Il est également possible de mémoriser le corpus sélectionné de façon à le réutiliser dans une session ultérieure.

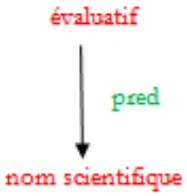
4.2 Recherche dans les textes

Une fois le corpus sélectionné selon les disciplines, les genres textuels et les parties textuelles désirés, l'utilisateur peut accéder au contenu du texte par trois types de recherche : un mode sémantique et guidé, axé sur la question linguistique du positionnement et du raisonnement, un mode de recherche libre et guidé, et un mode avancé utilisant des grammaires et des expressions régulières. Tous ces modes de recherche sont traduits dans le même langage de requête ConcQuest développé par Olivier Kraif (2008) et étendu par Achille Falaise (Falaise & Tutin 2010) dans le cadre du projet Scientext.

4.2.1 LE MODE S£MANTIQUE GUID£ : ACCÔS PAR LES GRAMMAIRES LOCALES

Le mode sémantique permet de faire des recherches dans les textes à partir de thèmes liés au positionnement ou au raisonnement, comme les opinions ou l'évaluation des objets scientifiques. Ce mode sémantique est construit à partir de schémas sémantico-rhétoriques, qui sont ensuite traduits dans le langage de requête, à l'aide de variables et de dépendances syntaxiques (Cf. Tutin 2010c ; Falaise & Tutin 2010).

A titre d'exemple, voici le schéma simple utilisé pour l'évaluation adjectivale des objets scientifiques, qui associe un élément évaluatif à un nom scientifique.



On fera correspondre à ce schéma une grammaire mobilisant pour chaque notion un ensemble de lexèmes (par exemple, pour les noms scientifiques : *méthode, problème, question,* etc. Pour les adjectifs évaluatifs, *adapté, pertinent, original, novateur ...*), alors que la relation pred (prédicat) sera traduite par la relation syntaxique épithète ou attribut. La syntaxe des grammaires, élaborée par Achille Falaise et Olivier Kraif, permet de redéfinir des relations, d'utiliser des variables et des raccourcis d'écriture. La principale difficulté est cependant de traduire des relations syntaxiques de surface en relations sémantiques, ce qui exige une excellente connaissance de l'analyse syntaxique de surface réalisée par le système Syntex (Bourigault 2007).

```
//TITRE: Adjectifs d'évaluation
//INFO: Les adjectifs d'évaluation qui portent sur les noms scientifiques
(ATTRIB,#2,#1) = (SUJ,#3,#1) (ATTS,#3,#2) ;
$eval=acceptable,adéquat,aisé,ambitieux,approximatif,bon,clair,classique,cohérent,complexe,
concis,confus,convaincant,correct,crucial,déterminant,difficile,discutable,encourageant,épineu
x,essentiel,excellent,faible,fin,flou,fondamental,important,innovant,intéressant,irréprochable,ju
dicieux,majeur,mauvais,meilleur,important,pertinent,nouveau,original,passable,passionnant,per
formant,principal,prometteur,riche,rigoureux,satisfaisant,séduisant,sérieux,significatif,solide,so
uhaitable,stimulant,vague,valable
$theo=
analyse,approche,article,caractéristique,choix,communication,concept,contribution,critère,élém
ent,étude,exemple,facteur,fonction,idée,méthode,modèle,notion,objectif,phénomène,problème,
projet,proposition,qualité,question,réflexion,résultat,solution,test,théorie,travail
Main = <lemma=$eval,#1> && <lemma=$theo,#2> :: (ATTRIB,#1,#2) OR (ADJ,#1,#2);
```

Ces grammaires sont ensuite intégrées dans l'interface et peuvent être librement choisies par l'utilisateur, comme cela apparaît sur la figure 5.

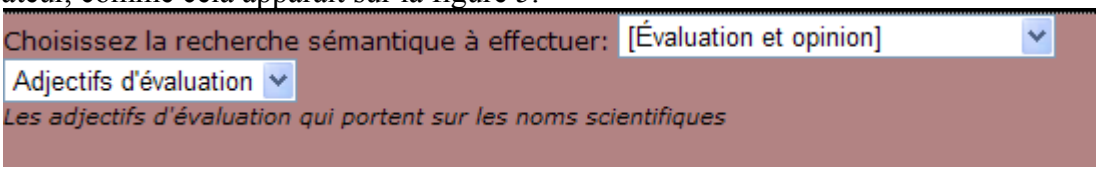


Figure 5 : Choix des requêtes sémantiques

Un ensemble de grammaires a été élaboré à cette fin autour de différents thèmes : évaluation, opinion, démarcation, auteurs cités, formulation des hypothèses, etc.

4.2.2 LE MODE SIMPLE ET GUID£

Ce mode de recherche guidé permet à l'utilisateur de sélectionner des formes, lemmes et/ou catégories, ainsi que les relations syntaxiques désirées. La requête présentée à la figure 6 permet ainsi d'extraire les occurrences où *hypothèse(s)* est le complément d'objet direct d'un verbe, indépendamment de la position des éléments dans la phrase. Cette requête extraira des exemples comme *faire des hypothèses, formuler des hypothèses, confirmer cette hypothèse ...* (Cf. Figure 8).

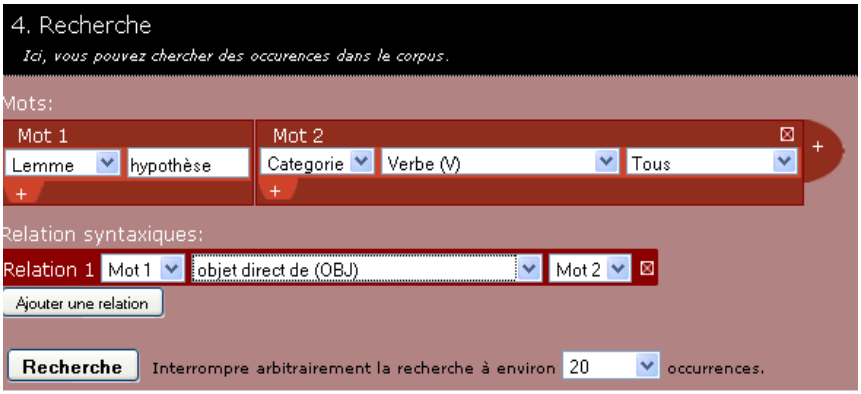


Figure 6 : Un exemple de requête libre et guidée (le nom hypothèse objet direct d'un verbe)

4.2.3 LE MODE AVANC£

L'utilisateur à l'aise avec les expressions régulières et la syntaxe de dépendance pourra aussi développer ses propres grammaires. Les grammaires exploitent la linéarité et/ou les dépendances syntaxiques. Elles permettent de redéfinir des relations syntaxiques, afin d'en proposer un traitement plus sémantique. Par exemple, dans la Figure 7, on définit une nouvelle relation SUJCOMP (sujet dans le cas des verbes composés) qui peut apparaître entre le sujet et le participe passé, par exemple dans *nous avons proposé*[20]. Les grammaires intègrent également des variables pour réaliser des raccourcis d'écriture.

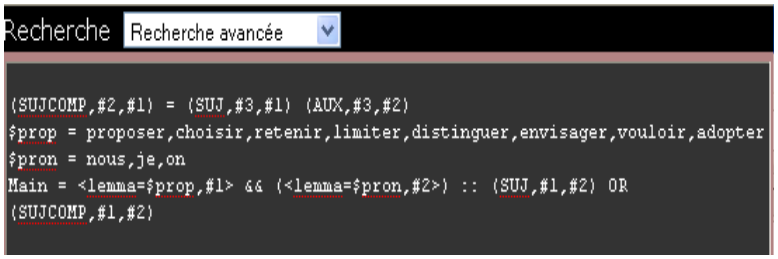


Figure 7 : Un exemple de recherche en mode avancé

4.3 Affichage et statistiques

Une fois la recherche effectuée, l'utilisateur peut ensuite faire afficher les résultats de plusieurs façons, les exporter afin de les traiter localement ou obtenir des traitements statistiques simples.

4.3.1 AFFICHAGE DES RÉSULTATS ET EXPORTATION

Après avoir effectué la recherche, le résultat s'affiche dans un concordancier KWIC, dont les fenêtres sont paramétrables. La référence du texte, ainsi que la partie textuelle, sont indiquées. L'utilisateur peut également demander un contexte plus large (dans la limite de 200 mots), comme dans la figure 8.

---

[20] Syntex est un analyseur syntaxique de surface. Dans *nous avons proposé*, il y a ainsi une relation SUJ entre l'auxiliaire et le sujet, et une relation AUX entre l'auxiliaire et le participe passé. A l'aide des redéfinitions de relations, on pourra ainsi proposer des analyses plus sémantiques du corpus de façon à analyser la relation profonde entre *nous* et *proposer* comme dans les exemples suivants : *nous avons pu proposer, nous venons de proposer, nous avons été contraints de proposer ….*

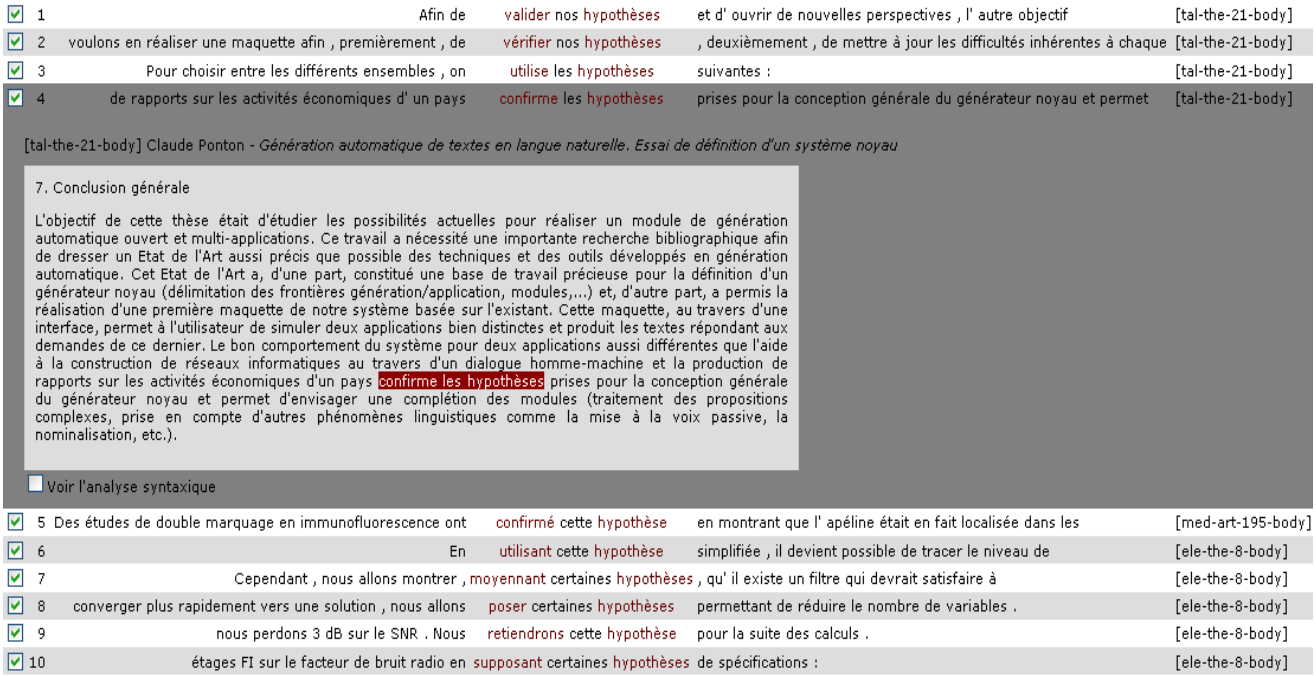


Figure 8 : Affichage des concordances : concordance KWIC et concordance large

On peut aussi obtenir l'affichage des structures syntaxiques, comme cela apparaît dans la figure 1.

Dans l'affichage KWIC, chaque occurrence peut être désactivée. Cela est particulièrement utile pour neutraliser les occurrences erronées du fait d'erreurs d'analyse syntaxique ou d'ambiguïté sémantique.

Enfin, l'utilisateur peut sauvegarder ses résultats, après avoir décoché les résultats inappropriés, dans un fichier HTML ou un fichier CSV (utilisable dans Excel), sur lequel il pourra retravailler ultérieurement.

4.3.2 STATISTIQUES

Des statistiques simples sur les résultats sont également intégrées à l'interface. On peut ainsi obtenir la liste des lemmes correspondant à une requête. La recherche sur les verbes ayant le lemme *hypothèse* comme objet direct (967 occurrences pour la totalité du corpus) montre ainsi que la collocation *faire* DET *hypothèse* est de loin l'expression la plus courante (Cf. Figure 9).

Au total 967 occurrences ont été trouvées.

Quelles statistiques souhaitez-vous consulter ?

Liste des lemmes

Lemme	Nombre	Forme
/faire/ /hypothèse/	242 (25.03%)	faire hypothèse (80) , faisons hypothèse (67) , fait hypothèse (45) , faisant hypothèse (12) , faire hypothèses (9) , font hypothèse (8) , ferons hypothèse (6) , faisions hypothèse (4) , faisons hypothèses (2) , Faites hypothèses (2) , fait hypothèses (1) , fasse hypothèse (1) , fais hypothèses (1) , fais hypothèse (1) , fit hypothèse (1) , faites hypothèses (1) , faisais hypothèse (1)
/tester/ /hypothèse/	80 (8.27%)	tester hypothèse (42) , tester hypothèses (17) , testé hypothèse (5) , testant hypothèses (3) , testerons hypothèse (3) , testons hypothèse (2) , testant hypothèse (2) , teste hypothèses (2) , teste hypothèse (1) , testent hypothèses (1) , testé hypothèses (1) , Tester hypothèse (1)
/confirmer/ /hypothèse/	64 (6.62%)	confirmer hypothèse (21) , confirment hypothèse (18) , confirme hypothèse (10) , confirmer hypothèses (4) , confirmé hypothèse (4) , confirment hypothèses (3) , confirmant hypothèse (2) , confirme hypothèses (1) , confirmerait hypothèse (1)
/formuler/ /hypothèse/	55 (5.69%)	formuler hypothèse (21) , formuler hypothèses (13) , formulons hypothèse (11) , formulé hypothèse (6) , formule hypothèse (2) , formulent hypothèses (1) , formulent hypothèse (1)
/émettre/ /hypothèse/	55 (5.69%)	émettre hypothèse (22) , émis hypothèse (8) , émettre hypothèses (6) , émettons hypothèse (5) , émet hypothèse (4) , émettent hypothèse (4) , émettons hypothèses (2) , émettrons hypothèse (1) , émis hypothèses (1) , émettent hypothèses (1) , émettant hypothèse (1)
/vérifier/ /hypothèse/	52 (5.38%)	vérifier hypothèse (37) , vérifier hypothèses (4) , vérifié hypothèse (3) , vérifient hypothèse (2) , vérifie hypothèse (2) , vérifient hypothèses (1) , vérifie hypothèses (1) , vérifions hypothèse (1) , vérifierons hypothèses (1)
/valider/ /hypothèse/	33 (3.41%)	valider hypothèse (14) , valider hypothèses (7) , valide hypothèse (4) , valident hypothèse (3) , validant hypothèse (2) , valident hypothèses (1) , validant hypothèses (1) , valide hypothèses (1)
/poser/ /hypothèse/	33 (3.41%)	poser hypothèse (8) , posons hypothèse (7) , poser hypothèses (7) , poserons hypothèses (2) , posé hypothèse (2) , poserons hypothèse (1) , posions hypothèse (1) , pose hypothèses (1) , posons hypothèses (1) , pose hypothèse (1) , posant hypothèse (1) , posé hypothèses (1)
/proposer/ /hypothèse/	28 (2.90%)	proposer hypothèses (9) , proposer hypothèse (7) , proposerons hypothèses (3) , proposé hypothèse (2) , propose hypothèse (2) , proposé hypothèses (1) , propose hypothèses (1) , proposent hypothèse (1) , proposant hypothèses (1) , proposons hypothèses (1)

Figure 9 : Les structures V-OBJ-> hypothèse les plus fréquentes

Il est également possible d'obtenir la répartition de la réponse dans les disciplines, les genres textuels ou les parties textuelles, en obtenant les fréquences absolues ou les fréquences relatives[21]. Ainsi, la grammaire des verbes d'opinion appliquée à la totalité du corpus (Cf. Figure 10) montre que ce type de verbe est fort fréquent dans les remerciements et les conclusions, mais moins usuel dans les introductions, notes et résumés (ce sont les fréquences relatives qui sont ici prises en considération).

[21] Fréquence relative : calcul du nombre d'occurrences sur le nombre total d'occurrences du texte.

Développement	1663	/	3645711	=	4.56 ‰
Introduction	69	/	209266	=	3.3 ‰
Conclusion	54	/	85200	=	6.34 ‰
Notes	49	/	153355	=	3.2 ‰
Annexe	24	/	118198	=	2.03 ‰
Remerciements	14	/	17551	=	7.98 ‰
Résumé	7	/	25061	=	2.79 ‰

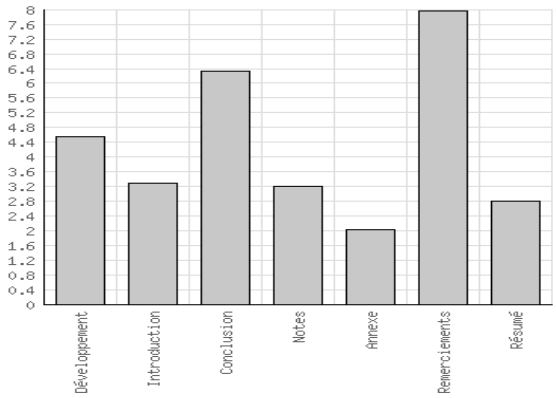


Figure 10 : Répartition des verbes d'opinion dans le corpus Scientext selon la partie textuelle

En ce qui concerne le genre textuel (Cf. Fig. 11), on observe d'intéressantes différences dans notre corpus. Si, de façon attendue, c'est dans les mémoires d'HDR que l'on trouve le plus d'expressions de l'opinion (verbale), on relève une intéressante différence entre les articles et les communications écrites qui sont souvent assimilés. La communication écrite serait-elle plus proche du genre oral, avec un engagement de l'auteur plus explicite ?

Au total 1557 occurrences ont été trouvées.

Quelles statistiques souhaitez-vous consulter ?

Répartition des lemmes

Genre	Nombre absolu d'occurrences		Nombre de mots total		Nombre relatif d'occurrences
Thèse	1146	/	3665882	=	3.13 ‰
HDR	188	/	454961	=	4.13 ‰
Communication	185	/	487188	=	3.8 ‰
Article	38	/	231903	=	1.64 ‰

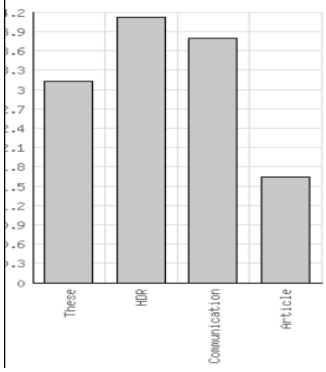


Figure 11 : Répartition des verbes d'opinion par genre textuel

Enfin, il est également possible d'obtenir la répartition des réponses par discipline.

Ces statistiques simples permettent de fournir rapidement des éléments sur la comparaison textuelle et disciplinaire. Ces résultats chiffrés ne doivent cependant pas être interprétés trop hâtivement et un retour au contexte sera souvent nécessaire pour désambiguïser une interprétation et analyser plus en finesse les formulations linguistiques.

5 CONCLUSION

Le projet Scientext est à la fois un projet ingénierique, avec l'élaboration d'un corpus d'écrits scientifiques diversifié librement mis à la disposition de la communauté linguistique à l'aide d'outils logiciels, et un projet de linguistique théorique visant à mieux comprendre le fonctionnement linguistique du positionnement et du raisonnement de l'auteur scientifique. Le premier volet est désormais réalisé, avec la mise à disposition d'un site web opérationnel, alors que le second volet, qui a fait l'objet d'un ensemble d'études sur la filiation, l'évaluation, la démarcation, les verbes de positionnement, le raisonnement causal, doit encore être développé. Ces premières études linguistiques permettent de brosser un portrait nuancé et diversifié de la question du positionnement de l'auteur, qui revêt plusieurs facettes qui ne sont pas nécessairement convergentes.

Nous espérons que le corpus Scientext qui, à notre connaissance, est un des seuls corpus analysés syntaxiquement librement consultables en ligne[22], sera largement exploité par la communauté des linguistes et sera suivi d'autres projets de ressources textuelles libres permettant de développer la linguistique de corpus en France.

6 RÉFÉRENCES

Boch F., Grossmann F. et Rinck F. (2009). « Le cadrage théorique dans l'article scientifique : Un lieu propice à la circulation des discours ». *Actes du colloque international Cit-dit*, Circulation des discours et liens sociaux: Le discours rapporté comme pratique sociale, Québec, du 5 au 7 octobre 2006, Nota Bene.

Bourigault D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Thèse d'habilitation à diriger des recherches. Université Toulouse le Mirail.

Cavalla C . et Tutin A. (à paraitre). « Etude des collocations évaluatives dans les écrits scientifiques ». Dans L. Gautier et S. Méjri (éds.). *Les collocations dans les discours spécialisés.* Dijon : Editions Universitaires de Dijon.

Chavez Ingrid (2008). *La démarcation dans les écrits scientifiques - Les collocations transdisciplinaires comme aide à l'écrit universitaire auprès des étudiants étrangers*. Mémoire de Master Français Langue Etrangère Recherche, ss.dir. Cristelle Cavalla. Grenoble : Université Stendhal-Grenoble3.

Falaise A. et Tutin A. (2010). Approche onomasiologique de la phraséologie transdisciplinaire des écrits scientifiques : la recherche sémantique dans les textes dans le cadre du projet Scientext, Démonstration, *Conférence THot*. 4-5 juin 2010, Annecy.

Florez M. (2010). « Marques de la citation positionnée dans trois disciplines des sciences humaines ». *Colloque international des Etudiants chercheurs en Didactique des Langues et en Linguistique*, du 29 juin au 2 juillet 2010, Université Stendhal, Grenoble, France.

Fløttum K., Dahl T. et Kinn T. (2006). *Academic Voices across languages and disciplines*. Amsterdam/Philadelphia : John Benjamins.

Garcia da Silva P. P. (2008). *Les marques de la filiation dans les écrits scientifiques*. Mémoire de Master 1, sous la direction de Francis Grossmann et d'Agnès Tutin, Université Stendhal-Grenoble.

Grossmann F., Tutin A. et Garcia da Silva P. (2009). « Filiation et transferts d'objets scientifiques dans les écrits de recherche ». *Pratiques* 143-144, p. 187-202.

Henderson A., Tutin A., Grossmann F. et Barr R. (2009). « SCIENTEXT : A Corpus of French and English Scientific Texts ». *British Association of Applied Linguistics Annual Conference*, 4

[22] Avec Corpus Eye, analysé avec le système VISL : http://beta.visl.sdu.dk/

septembre 2009, Newcastle University.
Hyland K. (2002). « Authority and invisibility : authorial identity in academic writing ». *Journal of Pragmatics*. Vol 34, 8, p. 1091-1112.
Hyland K. (1998*). Hedging in scientific research articles*. Amsterdam/Philadelphia : John Benjamins.
Kraif, O. et Tutin A. (2009). « Using a bilingual annotated corpus as a writing aid: An application for academic writing for EFL users. Dans N. Kübler (Éd.) *Proceedings of TaLC7, 7ème Conference Teaching and Language Corpora*. Bruxelles : Peter Lang.
Kraif O. (2008). « Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest ». *Actes des 9ème Journées d'analyse statistique des données textuelles, JADT 2008*. Lyon : Presses universitaires de Lyon. p. 625-634.
Nølke H., Fløttum K. et Norén C. (2004). *ScaPoLine. La théorie scandinave de la polyphonie linguistique.* Paris : Editions Kimé.
Rabatel A. (1998). *La construction textuelle du point de vue*. Lausanne/Paris : Delachaux et Niestlé.
Rinck F. (2006). *L'article de recherche en Sciences du Langage et en Lettres, Figure de l'auteur et approche disciplinaire du genre*. Thèse de Doctorat en Sciences du Langage, sous la direction de F. Boch et F. Grossmann, Université de Grenoble.
Rinck F., Boch F. et Grossmann F. (2007). *« Conformément à nos attentes*…, ou l'étude des marqueurs de convergence/divergence dans l'articlescientifique ». *Revue Française de Linguistique Appliquée,* XII-2, p. 109-122
Siddharthan A. et S. Teufel. (2007). « Whose idea was this, and why does it matter? Attributing scientific work to citations ». Dans *"Proceedings of NAACL/HLT-07*", Rochester, New York.
Tutin A. (coord.) (2007). « Lexique et écrits scientifique ». *Revue Française de Linguistique Appliquée*, volume XII-2, décembre 2007.
Tutin A. (2010a). Evaluative adjectives in academic writing in the humanities and social sciences. *Interpersonality in written academic discourse: perspectives across languages and cultures*. Cambridge : Cambridge Publishing. p. 219-239.
Tutin A. (2010b). « *Dans cet article, nous souhaitons montrer que* ... Lexique verbal et positionnement de l'auteur dans les articles en sciences humaines. Enonciation et rhétorique dans l'écrit scientifique ». *LIDIL 41,* p. 15-40.
Tutin A. (2010c). « Showing phraseology in context: an onomasiological access to lexico-grammatical patterns in corpora of French scientific writings ». Proceedings of eLexicography in the 21st century: new challenges, new applications, 22-24 october 2009, Louvain la Neuve.
Williams G. et Millon Ch. (à paraître 2010). « The General and the Specific : Collocational resonance of scientific language ». Proceedings Corpus Linguistics 2009. University of Liverpool

Remerciements

Nous remercions tout particulièrement D. Bourigault pour la mise à disposition de l'analyseur syntaxique Syntex. Un grand merci aussi à Kjersti Fløttum qui nous a autorisés à utiliser une partie du corpus KIAP pour cette étude.