La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral

Iris Eshkol-Taravella

La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral

MÉMOIRE

présenté pour l'obtention de

l'Habilitation à Diriger des Recherches

(spécialités : sciences du langage, traitement automatique des langues)

Iris ESHKOL-TARAVELLA

Composition du jury

Gabriel Bergounioux, PU, Université d'Orléans, LLL-CNRS
Catherine Schnedecker, PU, Université de Strasbourg, LiLPa (*rapporteur*)
Pierre Zweigenbaum, DR, LIMSI-CNRS (*rapporteur*)
Massimo Moneglia, PU, Université de Florence, LABLITA (*rapporteur*)
Isabelle Tellier, PU, Université Paris 3-Sorbonne Nouvelle
Denis Maurel, PU, Université de Tours, LI

Remerciements

Je voudrais remercier l'ensemble des membres du jury pour l'intérêt qu'ils ont manifesté à l'égard de mon travail. Je suis très reconnaissante à Catherine Schnedecker, Pierre Zweigenbaum et Massimo Moneglia d'avoir accepté d'être les rapporteurs de ce travail.

Je tiens à exprimer ma reconnaissance à Gabriel Bergounioux pour m'avoir accueillie au sein du laboratoire, pour sa disponibilité et pour son soutien tout au long de ce travail.

Je tiens à remercier Natalia Grabar et Céline Dugua pour leur relecture attentive et amicale et pour leurs remarques.

Le travail présenté dans cette HDR est aussi le produit de différentes collaborations menées au fil des ans avec Isabelle Tellier, Natalia Grabar, Catherine Domingues, Silvia Adler et Denis Maurel, dont la contribution professionnelle ne s'est jamais départie d'une relation de confiance.

Mes remerciements vont également aux membres du LLL et au département Sciences du Langage pour avoir su créer un environnement amical et stimulant dont j'ai largement bénéficié. Je remercie particulièrement Caroline Cance, Marie Skrovec, Céline Dugua, Lotfi Abouda, Emmanuel Schang, Flora Badin, François Nemo, Antonia Cristinoi, Maxime Lagrange, Linda Hriba et Layal Kanaan. Je suis profondément reconnaissante à Olivier Baude pour tout le travail entrepris sur la transcription de l'oral qui a mis au centre de mes études le corpus ESLO. Je remercie aussi mes étudiants pour leur confiance et nos échanges très stimulants pendant les cours.

J'aurai garde d'oublier mes amies, Emmanuelle Guerin, Catherine Lanoe, Corinne Laval, Nathalie Courtade, Hagar Mashari, Gal Kober, Olga Gofshtein et Alina Clément qui ont été à mes côtés au fur et à mesure de l'exécution de mon projet. Mes pensées vont aussi à Sarah Leroy qui m'a poussée à faire cette HDR. Sarah, merci. Je veux remercier enfin Aurélie Martin et Thomas Benatar pour les discussions passionnantes qui m'ont beaucoup stimulée dans ce travail.

En dernier lieu, je voudrais remercier toute ma famille d'ici et de l'étranger qui m'ont toujours apporté leur soutien : mon mari, mes parents, mes beaux-parents, Ella, Léa, Nina, Laurence, Galina et évidemment, ma fille, Tali, mon petit soleil d'amour.

Table des matières

1. Introduction

La rédaction d'un mémoire d'habilitation à diriger des recherches est l'occasion pour un enseignant-chercheur de faire le bilan de l'ensemble de ses activités de recherches. C'est l'occasion aussi d'avoir un regard réflexif et critique sur cette activité.

1.1 Parcours

Mon parcours universitaire se décompose en cinq périodes correspondant chacune à différents lieux, différentes personnes rencontrées et par là même différents disciplines.

- Ecole d'ingénieurs à Moscou : *informatique*

J'ai fait trois ans d'études à la Faculté d'automatique et d'informatique au Département des systèmes de mesure et de contrôle. Ce cursus m'a permis d'avoir une formation solide en informatique, mathématiques et électronique.

- Université de Tel-Aviv : *linguistique*

Après avoir émigré en Israël, j'ai poursuivi mes études à l'Université de Tel-Aviv où j'ai obtenu le diplôme de Licence de linguistique générale et de langue et littérature françaises. J'ai fait ensuite ma maîtrise et mon DEA en linguistique française. Cette formation m'a permis tout d'abord d'améliorer ma connaissance du français et de découvrir le domaine de la linguistique qui a décidé de la suite de mes activités de recherches. Mon mémoire, dirigé par David Gaatone, et intitulé *Comparaison de la structure [verbe support + nom prédicatif] avec le verbe simple correspondant. Le cas du verbe « donner »* a été mon premier travail de recherches, fondé sur un modèle inspiré de la méthode de Harris.

- Laboratoire de Linguistique Informatique (LLI) – Université Paris XIII : *description linguistique pour le TAL*

Mon travail de thèse *Typologie sémantique des prédicats de parole* a été réalisé sous la direction de Gaston Gross et Lucien Kupferman. Il s'inscrit dans le domaine de la linguistique appliquée. Il concerne la construction des classes homogènes de prédicats de parole en français. Deux modèles théoriques ont contribué à l'élaboration de cette recherche : d'une part, les *Classes d'objets et de prédicats* de Gaston Gross et d'autre part l'approche *Sens-Texte* développée par Igor Mel'čuk. L'objectif était d'utiliser ces classes dans les applications dans un Traitement Automatique du Langage (TAL) dédié à la traduction automatique et à l'extraction de'information.

- Laboratoire Langues, Logiques, Informatique et Cognition (LaLIC) - Université Paris IV : *TAL*

Afin d'acquérir une compétence concernant les applications des Sciences du Langage, j'ai fait parallèlement à ma dernière année de thèse un DESS *Ingénierie de la Langue et Société de l'Information*. J'ai pu mettre à profit les compétences acquises au cours de cette formation dans le cadre d'un stage portant sur l'extraction des entités nommées effectué chez *TEMIS*, une entreprise spécialisée dans le « Text Mining ».

- Laboratoire Ligérien de Linguistique (LLL) – Université d'Orléans

Ma nomination à Orléans et mon rattachement au CORAL, devenu le LLL, ont marqué un tournant dans ma carrière. Le travail sur corpus, en particulier les corpus oraux, constitue l'axe des recherches conduites au sein du laboratoire devenu l'UMR 7270 en 2012 avec quatre tutelles, les universités d'Orléans et Tours, la BnF et le CNRS.

Les études concernent principalement le développement et l'exploitation du corpus ESLO, collecté et transcrit par le LLL dans le cadre du projet ANR Variling et qui se poursuit dans le cadre de l'Equipex Ortolang. L'approche en linguistique variationniste est dominante dans la prise en considération des phénomènes. Ainsi, j'ai pu me familiariser avec la sociolinguistique, la linguistique du corpus et plus généralement l'ensemble des problématiques liées à la constitution, au traitement et à l'analyse du corpus oral.

Cette succession d'expériences et ce parcours de formation pluridisciplinaire m'ont permis d'acquérir des compétences dans des disciplines, sur des thèmes et avec formalismes variés qui m'ont orienté dans les activités de recherches que j'ai conduites depuis une douzaine d'années.

Ma nomination à Orléans en tant que Maître de conférences m'a permis également de découvrir un autre aspect des fonctions d'enseignant-chercheur puisqu'il m'a été confié la coordination d'une formation pluridisciplinaire. Mon recrutement avait été décidé dans la perspective de création d'une formation professionnalisante en TAL au niveau du Master à l'intérieur du département *Sciences du Langage*. Dès ma prise de fonction en 2003, j'ai participé à l'élaboration des maquettes et à la mise en place de cette spécialité du Master intitulé *Ingénierie Linguistique et Traitement de la Communication* (*ILTC*) renommé par la suite *Linguistique appliquée aux Sciences et Technologies de l'Information et de la Communication* (*LASTIC*), ainsi que du parcours *Communication et Traitement de l'Information* (*COMTIL*) en Licence qui en est la propédeutique. En tant que responsable de cette formation, il m'est demandé de définir le contenu des enseignements, de recruter des intervenants, d'encadrer les stages etc. Cette responsabilité m'a permis de mieux comprendre les enjeux, les possibilités et les difficultés d'une telle formation pluridisciplinaire qui associe la linguistique, le TAL et les métiers de la communication. En particulier, j'ai été conduite à me poser la question qui se trouve au centre de ce mémoire d'habilitation : Quel est le rôle du linguiste, et plus largement de la lingustique, dans le déploiement du TAL ?

1.2 Linguistique et TAL

1.2.1. Historique

L'origine du TAL peut être située aux États-Unis, où sont nées les premières idées de traduction automatique (TA) associées à l'apparition des machines électroniques. Les acteurs de la TA, aux États-Unis comme en France, sont des ingénieurs, des mathématiciens appliqués, des philosophes, des spécialistes de langues naturelles. Les linguistes sont au départ peu présents dans ces travaux.

Suite au rapport critique de (Bar-Hillel 1960) que confirme le rapport de l'ALPAC (Automatic Language Processing Advisory Committee) en 1966, le projet d'une traduction entièrement automatisée est sérieusement mis en doute et l'accent est mis sur une heuristique attendue des interactions entre langages formels logico-mathématiques, analyse grammaticale et programmation. Le développement des langages formels permet de reconsidérer les problèmes selon une approche déclarative où sont distingués la grammaire (la description linguistique) et les langages formels (qui rendent les informations linguistiques traitables par les ordinateurs). Les recherches privilégient les aspects psycholinguistiques de la syntaxe et de la sémantique aux USA et la linguistique algébrique dans des pays de l'Est. Dans cette perspective, l'automatisation du langage se trouve impliqué dans les investigations concernant l'Intelligence artificielle (IA). Ce rapprochement est rendu patent par les systèmes de représentation des connaissances et les trames (*frames*). Ainsi, depuis la fin des années 60, on assiste à la prédominance de modèles symboliques, au développement des grammaires

formelles et à l'émergence du champ de l'IA. Tous ces modèles utilisent des connaissances prédéfinies sur le monde et sur la langue pour construire les règles nécessaires au fonctionnement des systèmes.

Le début des années 90 marque un tournant dans les recherches du TAL. C'est une conséquence de la disponibilité d'un volume croissant de données linguistiques (corpus) au format numérique. Comme en témoigne l'un des pionniers de ces études : « […] la recherche basée sur corpus a vraiment décollé, non seulement comme un paradigme d'investigation linguistique reconnu mais comme une contribution clé pour le développement de logiciels de traitement du langage naturel. La recherche […] va probablement susciter non seulement l'attention des universitaires mais le financement industriel et public qui sera nécessaire si l'on veut obtenir les progrès souhaités. » (Leech 1991 : 20).

1.2.2. Succès des méthodes statistiques

Le corpus est devenu « source de connaissances » pour l'élaboration de ressources lexicales telles que les dictionnaires, les thésaurus ou les ontologies et « objet d'étude » pour l'analyse par des outils informatiques (Nazarenko 2006). (Cori *et al.* 2008 : 6-7) distinguent cinq types d'usages des corpus :

- la mise à disposition versatile des corpus pour la communauté ;

- l'élaboration, grâce aux corpus, des outils linguistiques comme bases de données, dictionnaires, grammaires etc. ;

- les descriptions linguistiques de formes à partir de leur usage en contexte ;

- le traitement de la variation ;

- la construction d'outils de TAL à base de corpus d'entraînement.

Cette diversité d'usages a conduit à l'emploi croissant de méthodes statistiques qui permettent un traitement rapide et de généricité maximale, c'est-à-dire un traitement adapté à n'importe quel corpus. (Nazarenko 2006) note que si les travaux actuels en TAL privilégient l'extension des corpus plutôt que l'exhaustivité des phénomènes, cela tient entre autres à ce que les méthodes statistiques requièrent des volumes importants de données. Au nombre des méthodes statistiques, on peut citer les techniques d'apprentissage automatique ou la statistique textuelle. L'apprentissage automatique supervisé est devenu une méthode très répandue dans les tâches de l'annotation des corpus ou encore de la classification des textes. La statistique textuelle a fourni, par exemple, des résultats incontestables dans l'étude des collocations ou la définition des genres textuels. Les connaissances préalables des experts du domaine et du linguiste ne sont plus impérativement nécessaires dès lors qu'elles peuvent être reconstitées et acquises directement à partir des données traitées. La prédominance des méthodes quantitatives est aujourd'hui patente dans le TAL où tout se mesure en chiffres : le corpus, l'évaluation de la méthode et le résultat.

1.2.3. Linguistique outillée

Les possibilités offertes par le TAL et notamment les techniques d'exploitation des documents numériques ont permis des développements théoriques fondés sur l'exploitation de corpus, mettant ceux-ci aux centres de la description et de l'analyse linguistiques. Elles sont devenues l'atout principal de la linguistique de corpus. Des outils comme Lexico, TXM, Hyperbase etc. peuvent désormais être utilisés par les linguistes.

Le TAL se situe ainsi comme un auxiliaire, une technique ou, comme la désigne (Habert 2004, 2006), « un instrument » qui permet aux linguistes de tester leurs hypothèses ou de vérifier leurs théories. Habert oppose des logiciels (étiqueteur, concordancier ou logiciel d'aide à la transcription de l'oral) dédiés à un traitement automatique des données linguistiques, qu'il définit comme des « instruments », et des logiciels multi-usage comme le tableur Excel ou un gestionnaire de base de données référentielle qu'il appelle des « outils ». Dans cette perspective, les chercheurs en TAL développent des « instruments » pour les linguistes qui les utilisent pour l'analyse des concordances, des calculs de fréquence de mots, les recherches lexicographiques, la confection de dictionnaire à consultation automatique etc.

1.2.4. Une pluridisciplinarité problématique

Le TAL est aujourd'hui devenu un domaine autonome qui se situe au carrefour de trois disciplines : linguistique, informatique et mathématiques. Pourtant, les articles qui traitent du TAL sont le plus souvent publiés dans des revues d'informatique. Inversement, peu d'articles de TAL paraissent dans des revues généralistes de linguistique ou de mathématiques.

En 1993, le Ministère a défini un nouvel intitulé dans la nomenclature des diplômes nationaux : la licence de Sciences du langage, mention « Traitement automatique des langues », que prolonge la maîtrise en Sciences du langage, mention « Industries de la langue ». Le TAL se trouve donc inscrit comme une application des sciences du langage et un étudiant en SDL qui se spécialise en TAL se doit d'acquérir des compétences en informatique (langages de programmation, technologies du Web, gestion des bases de données etc.), ses compétences linguistiques étant déjà assurées. Sur le marché industriel ce sont surtout les compétences en informatique qui sont prisées afin d'assurer le développement de logiciels dédiés au TAL.

Ainsi, paradoxalement, même si les formations en TAL sont proposées au sein des départements SDL, la linguistique semble de moins en moins présente dans le TAL qui privilégie les applications industrielles au détriment des investigations théoriques. On constate « deux lignes de tension constantes dans l'histoire du TAL : la cohabitation paradoxale et nécessaire des recherches théoriques et des applications à visée industrielle d'une part, les antagonismes entre le TAL et les différentes disciplines qui le constituent, voire entre ces disciplines elles-mêmes quand elles rentrent en interaction dans un problème de TAL ». (Cori et Léon 2002).

1.3 Mon positionnement scientifique

Mes travaux de recherches s'inscrivent dans le domaine du TAL. Depuis dix ans, je mets à profit mes connaissances linguistiques pour améliorer des solutions informatiques. Plusieurs difficultés ont aiguillé mes recherches. Etant de formation linguistique, j'ai dû considérer ma situation de « taliste », celle d'une *taliste* pour les linguistes, celle d'une *linguiste* pour les talistes.

De mon point de vue, le linguiste peut avoir deux types de relations aux « outils » ou aux « instruments » informatiques :

- soit être un simple utilisateur des « outils » et des « instruments » : dans ce cas, il intervient en linguistique de corpus ou en linguistique outillée ;

- soit contribuer au développement des « instruments » et se place en tant qu'acteur du traitement automatique au même titre qu'un informaticien.

Si le premier rôle est bien avéré, le second reste souvent mal défini.

Ce mémoire m'offre l'occasion de revenir sur le rôle qu'une linguiste-taliste peut jouer dans ce domaine et sur la façon dont la linguistique peut contribuer aux travaux et aux résultats du TAL.

Le spécialiste en TAL aujourd'hui ne peut pas se contenter de compétences informatiques. Il doit être à même de constituer un corpus selon les méthodes et les techniques actuelles, connaître les formalismes et les modèles utilisés, savoir analyser les résultats obtenus au-delà de leur quantification statistique. Ce travail demande une grande rigueur et des capacités d'observation des variations et des régularités attestées par le corpus pour assister, perfectionner ou interpréter le processus automatique. Ce sont ces compétences que j'ai essayé de transmettre à mes étudiants de Licence et de Master à travers les cours « Constitutions de corpus, Outils linguistiques pour l'extraction de l'information, Enrichissement des corpus, Description linguistique pour le TAL, Traitement de l'information etc. » Ces principes ont également été mis en œuvre dans mes recherches qui concernent le repérage et l'analyse d'une information linguistique dans les corpus.

Deux préoccupations ont paramétré mes travaux: la prise en considération de la nature des corpus traités et la modélisation de l'information linguistique destinée à l'analyse. La nature particulière des données (il s'agit de corpus « non standards » : corpus oral sociolinguistique ou corpus des titres de cartes géographiques issus du Web) et de l'information recherchée (renseignements sur le locuteur, reformulations paraphrastiques, actions et commentaires dans les recettes de cuisine etc.) rend difficile l'application de processus automatiques. Les outils pour traiter ce genre de corpus et ce type d'information sont rares et/ou inacessibles. La méthodologie adoptée, commune à tous mes travaux, suit quatre étapes :

1. Analyse manuelle du corpus : l'objectif de cette étape est de se familiariser avec les données traitées et d'observer les variations dans les occurrences.
2. Modélisation : l'analyse préalable des données permet de modéliser l'information qu'on cherche à repérer. Cela consiste à établir une typologie sous forme d'un jeu d'étiquettes correspondant à la nature du corpus, c'est-à-dire en tenant compte de ses spécificités d'une part et de l'objectif assigné d'autre part.
3. Etablissement de la technologie adaptée : au moment de la prise de décision sur le choix de la technologie, le choix est déterminé par le respect des données linguistiques et des contraintes que ces données imposent au traitement automatique ainsi que par la finalité déclarée. Il s'agit souvent d'opter entre le développement d'un nouvel outil ou bien une adaptation d'un outil existant.
4. Analyse quantitative et qualitative des résultats.

Ma démarche est résolument empirique, guidée par les observables issus des corpus. Elle est fondée sur la préservation des spécificités linguistiques des corpus traités et sur la prise en compte de la variation linguistique présente dans ces corpus. Je considère que le travail préalable au traitement et à l'exploitation du corpus constitue, au même titre que les résultats quantitatifs, un apport appréciable en TAL. Constitution des données et traitement sont indissociables et doivent être suivis de la collecte jusqu'à la diffusion du corpus. Les outils ne sont pas préexistants aux données. Ils doivent répondre aux besoins particuliers liés aux corpus à traiter et/ou analyser.

Le mémoire se compose de grandes parties en adéquation avec la nature du corpus traité. Tout d'abord, mes travaux sur l'annotation du corpus oral. Il s'agit :

- du traitement de corpus et plus précisément de la préparation requise pour en assurer l'exploitation optimale par des chercheurs en linguistique entre autres (*annotation syntaxique* et *anonymisation*) ;

- du repérage de l'information sémantique (*annotation de l'information personnelle sur le locuteur, annotation des reformulations paraphrastiques, annotation des lieux*).

- de l'analyse des données orales (*étude des commentaires et des actions dans le corpus des recettes de cuisine*).

La deuxième partie est consacrée à l'annotation et à l'analyse des désignations des lieux dans le corpus Web des titres de cartes géographiques. Il s'agit d'un corpus écrit non normalisé.

La troisième partie décrit mon travail sur un corpus plus standard, le corpus médiatique du *Monde*. Ce travail entre dans le domaine de la linguistique outillée et concerne l'étude des noms dits généraux.

Ce cheminement reflète une progression, depuis le traitement de corpus non standard et hors normes – transcriptions d'ESLO et corpus Web, dans une fluctuation entre l'oral et l'écrit –, sur lesquels les méthodes « classiques » du TAL éprouvent des difficultés jusqu'au corpus normalisé du *Monde* où sont étudiés les noms généraux. Ce phénomène, caractéristique de ce type de corpus, intervient à un tel degré d'abstraction et avec si peu homogénéité qu'il pose également des problèmes aux outils du TAL. Pour pallier ces difficultés, des connaissances et une expertise spécifiques doivent être mobilisées. C'est dans ce domaine que l'intervention du linguiste peut apporter une contribution non négligeable.

2. Annotation

2.1. Définition

Globalement, l'annotation consiste dans l'apport d'informations de nature différente. On parle à ce sujet d'une « valeur ajoutée » (Leech 1997) aux données brutes. Je distinguerai trois types d'annotations qui s'appliquent à trois domaines différents et à des applications distinctes :

- l'annotation dans son sens premier comme ajout manuel de remarques, commentaires, notes sur le texte ;

- l'annotation du document et/ou du corpus avec les métadonnées caractérisant et décrivant le document numérique ;

- l'annotation d'ordre linguistique dans le cas de l'étiquetage morphosyntaxique ou de l'annotation sémantique.

Les années 1990 ont constitué un tournant dans l'évolution du traitement automatique du langage (TAL) avec la constitution et l'exploitation de corpus qui ont provoqué une redéfinition des objectifs et un renouvellement des méthodes de la linguistique et du traitement automatique (Habert et Nazarenko 1997, Nazarenko 2006). Pour pouvoir accéder au contenu du corpus, le traiter et l'analyser, le processus de l'annotation est devenu indispensable.

À toutes les étapes et dès la collecte des données, se pose la question de l'annotation des métadonnées, qui sont des éléments descripteurs de la ressource afin de faciliter son exploitation, sa réutilisation et son archivage. Pour les corpus oraux, s'ajoute une étape préalable, celle de la transcription. Le processus de la transcription peut être considéré comme

un enrichissement de l'information sonore au moyen d'une information orthographique : à ce titre, la transcription peut être considérée comme une annotation. Elle est obligatoire pour permettre une exploitation des données orales, les outils informatiques ne permettant pas aujourd'hui de travailler directement sur le signal. L'enrichissement des occurrences par l'ajout d'une information grammaticale sur la catégorie syntaxique (POS), le genre, le nombre etc. est important pour la mise à disposition et la consultation des données car il permet de faire des requêtes précises à partir de ces indications. Les différents phénomènes linguistiques annotés sont directement accessibles aux chercheurs et permettent une analyse plus fine.

2.2. Outils d'annotation

Les outils d'annotation varient selon la nature de l'annotation, c'est-à-dire selon les phénomènes que l'on veut distinguer. Ainsi, l'annotation automatique des coréférences, par exemple, pose de nombreux problèmes et nécessite encore aujourd'hui le recours à l'intervention humaine (Mélanie-Becquet et Landragin 2014, Muzerelle *et al.* 2014). Toutefois, il existe des outils d'aide à l'annotation manuelle comme Transcriber[1], Praat[2], ANVIL[3], ELAN[4], Advene[5] pour la transcription des fichiers audio et vidéo, Glozz[6], Gate[7], MMAX2[8], Knowtator[9], Calisto[10], ANALEC[11] etc. pour d'autres niveaux d'annotation. Ils permettent de réduire l'effort nécessaire à la production de corpus annotés et de réaliser, parfois, diverses vérifications, en particulier pour ce qui concerne la cohérence.

L'annotation automatique ou semi-automatique peut se faire avec des méthodes à base de règles linguistiques décrivant le contexte d'emploi de phénomènes à annoter sous forme de grammaires locales ou avec des méthodes d'apprentissage automatique à partir d'un corpus de référence annoté manuellement. Les méthodes hybrides combinent les deux techiques.

2.3. Formats d'annotation

La constitution du corpus annoté pose le problème du format des données annotées. Il existe différentes normes et conventions sur l'annotation des données comme Ester ou Quaero pour les entités nommées, timeMl pour les expressions temporelles et événements, TEI pour le codage des métadonnées etc. Cependant, il n'est pas toujours possible d'être conforme à ces normes s'il s'agit d'un phénomène qui n'a pas été pris en compte dans les conventions proposées. Se pose alors la question d'adapter les étiquettes à celles normalisées ou de développer un nouveau jeu d'étiquettes qui permettra de mieux représenter le phénomène en question.

La sortie de l'annotation peut varier selon les outils et les méthodes appliquées, selon que la distinction des éléments s'effectue par des balises (XML, HTML etc.) ou des accolades (Unitex). On peut séparer le document en tokens et attribuer les étiquettes sous forme de colonnes (TreeTagger, SEM).

---

1. http://trans.sourceforge.net/en/presentation.php. Une nouvelle version de ce dernier est disponible depuis juillet 2011
2. http://www.fon.hum.uva.nl/praat/
3. http://www.anvil-software.de/
4. http://icar.univ-lyon2.fr/projets/corinte/confection/elan.htm
5. http://liris.cnrs.fr/advene/
6. http://www.glozz.org/
7. http://gate.ac.uk/
8. http://mmax2.sourceforge.net/
9. http://knowtator.sourceforge.net/index.shtml
10. http://callisto.mitre.org/
11. http://www.lattice.cnrs.fr/Analec,68

2.4. Processus subjectif d'annotation

En définissant l'annotation comme une « valeur ajoutée » consistant en un apport d'informations de nature *interprétative* aux données brutes, (Leech 1997) compare ce processus avec l'interprétation, introduisant un caractère subjectif qui s'exprime à travers la sélection des données à annoter.

L'annotation est une façon de s'approprier le corpus. Les différents annotateurs humains interprètent et perçoivent différemment les données. Les résultats d'annotation peuvent dépendre non seulement de leurs connaissances du domaine annoté et de leur orientation théorique mais aussi de variables sociologiques. Le guide d'annotation se doit donc d'être le plus clair, le plus exhaustif et le moins ambigu possible. C'est ce que notent aussi (Mélanie-Becquet et Landragin 2014) : « Pour que les annotations ne soient pas trop subjectives, un manuel d'annotation strict et directif s'avère nécessaire. Il faut cependant que le schéma d'annotation tienne compte des ambiguïtés et flous possibles, et autorise une certaine souplesse dans l'affectation des valeurs. » Un moyen d'objectiver les résultats consiste dans les calculs d'accords inter- et intra-annotateur qui servent à quantifier la fiabilité, et donc la qualité, des annotations produites. Les mesures Kappa (κ) de Cohen (Cohen 1960) et de Carletta (Carletta 1996) normalisent l'accord observé en fonction de l'accord attendu (ou dû au hasard).

Dans le cas de l'annotation automatique, les annotations diffèrent aussi selon les outils, c'est-à-dire les choix théoriques et méthodologiques faits en amont par les concepteurs des logiciels.

Le jeu d'étiquettes n'est jamais universel. Il dépend directement de l'école, du modèle théorique dans lequel s'inscrit l'annotateur. Il ne peut jamais être exhaustif. Le nombre et le contenu des étiquettes peuvent varier d'un outil à l'autre. Même des étiquettes identiques peuvent avoir une extension très différente d'un système à l'autre. Certains manques et certaines imperfections apparaissent comme inhérents. Il faut également prévoir certains aménagements si l'on cherche à automatiser le processus. Le nombre d'étiquettes peut être ainsi réduit pour perfectionner le système d'annotation automatique par apprentissage, par exemple. « Il n'y pas de meilleur jeu d'étiquettes, […] dans la pratique la plupart des jeux d'étiquettes constituent plutôt des compromis entre la finesse de la description linguistique et ce qui peut être attendu, pour des raisons pratiques, d'un système automatique d'étiquetage » (Leech 1994 : 51).

Ainsi, il n'y a pas une seule version de corpus annoté mais plusieurs - existantes ou potentielles.

2.5. Conclusion

Le corpus annoté peut être considéré comme une nouvelle version du corpus d'origine. La réflexion sur la méthodologie de sa constitution doit être instruite en fonction de la nature des données linguistiques à annoter, et d'autre part de l'utilisation finale. Tous les choix sur le jeu et format d'étiquettes, sur le contenu de l'information à annoter, sur l'outil et la technologie à utiliser, doivent intégrer ces deux aspects[12].

De mon point de vue, l'annotation n'est pas un processus exclusivement technique. Tout comme la constitution de corpus, elle soulève de nombreuses questions parmi lesquelles les

[12] Ces réflexions ont été menées, entre autres, dans le cadre du projet Ancor (2011-2013) financé par la région Centre et consacré à la création d'un corpus oral annoté en anaphores et coréférences (Muzerelle *et al.* 2013, 2014).

questions linguistiques occupent une place importante. C'est sur ce point que le rôle imparti au linguiste est déterminant.

Le processus d'annotation est aussi un processus subjectif. Le travail que j'ai pu effectuer sur l'annotation des données témoigne de la forme d'appropriation effectuée sur ces corpus. Les choix concernant la méthodologie et les étiquettes ont été dictés par la nature des données et, d'une manière implicite, par une certaine perception, qui est une forme d'interprétation du corpus.

3. Annotation de l'oral

Force est de constater que l'oral a été longtemps marginalisé dans le champ de la linguistique française (Blanche-Benveniste et Jeanjean 1987) comme dans celui de la linguistique de corpus. Faisant l'inventaire des corpus oraux en français, (Cappeau et Gadet 2007) notent qu' « il n'y a pas eu en France de volonté institutionnelle qui aurait conduit à la constitution d'un grand corpus oral. C'est, en contraste, ce qui a été fait pour l'écrit ». Cependant les travaux sur « le français parlé » puis l'apport des nouvelles technologies ont permis un engouement récent pour ce domaine. Parmi les initiatives actuelles, on peut citer la base CLAPI[13] constituée pour étudier les interactions orales, le corpus PFC[14] plus particulièrement consacré à l'analyse de certains phénomènes phonologiques, le corpus CRFP[15] pour la morphosyntaxe ou corpus de français spontané EPAC[16] composé d'interviews et de débats d'émissions télévisées.

Des initiatives institutionnelles (Centre de ressources numériques du CNRS, ANR Corpus, Programme Corpus de la parole de la DGLFLF en partenariat avec les fédérations de recherche en linguistique du CNRS, la création du TGE-ADONIS dont l'objectif était de mutualiser les ressources, standards technologiques et préserver des données dans les SHS en collaboration avec le réseau des centres de gestion de ressources et de technologies linguistiques CLARIN et de la TGIR Huma-Num) rendent possibles la mise à disposition de corpus oraux d'envergure.

Pour exploiter un corpus oral, il est nécessaire de le transcrire et certaines tâches d'annotation deviennent dès cette étape utiles et/ou indispensables. Les choix d'annotation diffèrent d'un projet à l'autre suivant des objectifs variés. Ainsi, dans le cadre du projet OTIM[17], le travail d'annotation a porté sur un grand nombre de domaines : phonétique, prosodie, phonologie, syntaxe, discours et gestes. Le corpus EPAC a été annoté en prenant en compte divers phénomènes : bruits, musiques, inspirations, prononciations particulières ou erronées, mots étrangers, néologismes... Le projet ANR Rhapsodie[18], quant à lui, a mis au centre de ses activités les annotations prosodique et syntaxique des données orales existantes. La suite de Rhapsodie, le projet ANR Orfeo (2012-2016) propose la constitution d'un Corpus d'Etude pour le Français Contemporain (CEFC) annoté entre autre par les informations morphologiques, syntaxiques, macro-syntaxiques, sémantiques, conversationnelles et prosodiques[19].

[13]. Corpus de langues parlées en interaction, http://clapi.univ-lyon2.fr/
[14]. Phonologie du français contemporain, http://www.projet-pfc.net/?accueil:intro
[15]. Corpus de référence du français parlé, http://www.up.univ-mrs.fr/delic/crfp
[16]. Exploration de masse de documents audio pour l'extraction et le traitement de la parole conversationnelle, http://projet-epac.univ-lemans.fr/doku.php?id=accueil.
[17] Outils pour le Traitement de l'Information Multimodale, http://www.lpl-aix.fr/otim
[18] http://rhapsodie.risc.cnrs.fr/fr/index.html
[19] http://www.lattice.cnrs.fr/ORFEO-Outils-et-Ressources-pour-le

3.1. Caractéristiques de l'oral

3.1.1. Disfluences

Le langage oral est différent de l'écrit du fait de phénomènes tels que les *disfluences.*

Pour (Blanche-Benveniste *et al.* 1990), il s'agit d'une accumulation d'éléments qui « brisent le déroulement syntagmatique » sans rien ajouter à la sémantique de l'énoncé. L'écrit se présente au destinataire comme un produit final alors que l'oral est un produit en cours d'élaboration. « […] le scripteur peut revenir sur ce qu'il a écrit, pour le corriger ou le compléter. A l'oral, […] toute erreur, tout raté ou mauvais départ ne peuvent être corrigés […] que par une reprise, une hésitation voire une rupture de construction qui laissent des traces dans le message même. » (Riegel *et al.* 1994 : 30). L'oral montre les traces de sa propre élaboration à la manière de brouillons qui précèdent la version finale de nos écrits (Blanche-Benveniste *et al*. 1990 : 17).

Pour (Dister 2007), les disfluences sont les « marques typiques des énoncés en cours d'élaboration » qui «constituent un piétinement sur l'axe syntagmatique de l'énoncé et […] nécessitent d'être prises en compte par le système d'étiquetage. »

Les disfluences constituent un problème pour l'analyse automatisée de l'oral (Adda-Decker *et al.* 2003, Antoine *et al.* 2003, Benzitoun 2004, Benzitoun *et al.* 2004, Valli et Véronis 1999 etc.) car elles réduisent considérablement les performances d'outils conçus pour de l'écrit standard. « Mais c'est certainement une erreur que d'imaginer que le modèle suivi pour l'écrit pourrait être transféré à l'oral. En effet, les corpus oraux sont liés à des exploitations extrêmement diversifiées (analyse prosodique, analyse de discours, analyse syntaxique, approches pragmatiques ou sociolinguistiques etc.) qui nécessitent des informations par nature très disparates. » (Cappeau et Gadet 2007)

Parmi les disfluences on retrouve :

- des hésitations : *madame **euh** comment vous faîtes une omelette*
- des faux-départs : *il va y avoir encore **des** encore mais*
- des répétitions : ***le le***
- autocorrections : *juste après **le la** fin du premier cycle*
- des reformulations : *on fait ce que l'on appelle un carton c'est-à-dire le le ce dessin-là agrandi*
- amorce : *vous **v-** vous, vous êtes **in-** institutrice*

etc.

(Dister 2007) regroupe sous le terme de disfluences : « les répétitions, les corrections directes, liées aux répétitions, les amorces de morphèmes, le morphème *euh*. » Cette typologie vise à décrire les données orales (le corpus Valibel) et est utilisée pour traiter les disfluences dans le cadre de l'étiquetage morphosyntaxique.

Les disfluences ont été étudiées et classées par l'action COPTE (Corpus Parole/ Texte et Évaluation) dans l'objectif d'améliorer la reconnaissance automatique de la parole. Cette classification a suivi les recommandations du Linguistic Data Consortium (LDC)[20] pour l'annotation des disfluences dont le guide de transcription (annotation) pour l'anglais oral avait pour but de rendre plus directement utilisables les transcriptions qui serviront à un traitement automatique, comme l'extraction ou l'alignement de données. COPTE distingue sept types de disfluences :

[20] https://www.ldc.upenn.edu/

- les « pauses remplies » (*euh*) ;
- les marqueurs discursifs (*disons, eh bien...*) ;
- les marques d'édition du locuteur concernant ses propres paroles (*il fait moche, enfin, je veux dire, il y a du vent et de la pluie*) ;
- les apartés (*cette question, qui par ailleurs est très amusante, m'embarrasse*) ;
- les répétitions (*le le*) ;
- les révisions (*le la*) ;
- les amorces (des mots interrompus en cours de réalisation)

Les disfluences doivent être prises en compte au cours du traitement automatique. Mes travaux sur le corpus oral constituent des exemples de ce type de traitement, et ce de plusieurs manières :

- dans le cas de l'annotation syntaxique, les étiquettes correspondantes ont été attribuées aux disfluences au même titre que les autres unités du discours (Eshkol *et al.* 2010, Tellier *et al.* 2013, 2014);
- les disfluences ont été prises en compte dans le cas du repérage automatique de l'information sémantique grâce aux règles établies (Maurel *et al.* 2011, Eshkol-Taravella *et al.* 2012) ;
- un type de disfluences a été annoté et étudié à part, à savoir des reformulations paraphrastiques (Eshkol-Taravella et Grabar 2014a,b, Grabar et Eshkol-Taravella 2015).

3.1.2. Transcription

A la différence de l'écrit, un corpus oral associe parole collectée et transcription. « On ne peut pas étudier l'oral par l'oral, en se fiant à la mémoire qu'on en garde. On ne peut pas, sans le secours de la représentation visuelle, parcourir l'oral en tous sens et en comparer les morceaux. » (Blanche-Benveniste 2000 : 24). Ce paradoxe a été mis en évidence par de nombreux chercheurs (Blanche-Benveniste et Jeanjean 1987, Blanche-Benveniste 1997, 2000, Gadet 2003, Raingeard et Lorscheider 1977) qui constatent que pour approcher l'oral, on doit « en passer » par l'écrit, c'est-à-dire par sa transcription.

Les transcriptions ne sont en général pas ponctuées pour éviter l'anticipation de l'interprétation (Blanche-Benveniste et Jeanjean 1987). Selon les auteurs, en ponctuant, le transcripteur « suggère une analyse avant de l'avoir faite » (1987: 142).

De même la notion de phrase, essentiellement graphique, a rapidement été abandonnée par les linguistes qui s'intéressent à l'oral. Gadet (1992 : 69) note que « pour toutes les études de phénomènes oraux, la séquence fondamentale ne correspond généralement pas à ce que l'on entend par « phrase » à l'écrit. Il faut donc se passer de cette catégorie ».

Les transcriptions consignent les marques du travail de formulation, les disfluences.

Suite aux travaux en linguistique de l'oral, je partage l'avis qu'on ne peut pas traiter de la même manière l'écrit et l'oral. Même sur un genre bien défini et contraint comme une recette de cuisine, la comparaison entre les recettes écrites (dans un manuel de cuisine ou sur les sites Web) et celles proposées à l'oral accusent la différence (Bergounioux et Eshkol, à paraître). La situation de communication, la présence d'un interlocuteur, les cadres du dialogue, la personnalité du locuteur, la perception de la question etc. sans compter les caractéristiques propres de l'oral, introduisent des variations que l'écrit ne présenterait pas. Le traitement et l'analyse de l'oral sont l'une des caractéristiques des travaux d'annotation que j'ai effectués sur le corpus ESLO (Enquêtes Sociolinguistique à Orléans) du LLL.

3.2. Corpus ESLO

La première Enquête SocioLinguistique à Orléans, ESLO1, a été conçue il y a quarante ans dans une perspective de didactique en Français Langue Etrangère. Ce sont des enregistrements recueillis par des chercheurs britanniques auprès de différents groupes de la population orléanaise dans les années 1968-1969. ESLO1 « comprend environ 200 interviews, toutes référencées (caractérisation sociologique des témoins, identification de l'enquêteur, date et lieu de passation de l'entretien) » (Abouda et Baude 2007:164), mais aussi une gamme d'enregistrements variés (des reprises de contacts informelles comme des discussions entre amis, des enregistrements en micro caché, des conversations téléphoniques, des réunions publiques, des transactions commerciales, des repas de famille, des interviews de personnalités de la ville (monde politique, syndical, universitaire ou religieux), des conférences ou débats ainsi que des entretiens au Centre Médico-Psychopédagogique d'Orléans (entretiens entre une assistante sociale et des parents). Le corpus représente 300 heures (environ 4 500 000 mots).

Dans les années 1980-90, une partie du corpus a été transcrite et étiquetée puis mise à disposition sur la toile dans le cadre du projet ELILAP/LANCOM[21]. Dans les années 1993-2001, le corpus a été repris par des chercheurs de l'Université de Louvain (Debrock *et al.*, 2000).

Dans le cadre du projet ANR Variling, la totalité d'ESLO1 a été transcrite et une nouvelle enquête ESLO2 a été entreprise en 2008. ESLO2 est un corpus en incrémentation continue. À terme, il comprendra plus de 350 heures d'enregistrements afin de former avec ESLO1 un corpus d'environ 700 heures et de dix millions de mots. Il s'agira alors d'un grand corpus oral réalisé selon des bonnes pratiques de constitution garantissant l'interopérabilité des données avec d'autres projets semblables[22].

Le travail d'annotation a été effectué sur les fichiers transcrits à l'aide de Transcriber.

Les conventions de transcription respectent deux principes : l'adoption de l'orthographe standard et le non-recours à la ponctuation de l'écrit. Les marques typographiques comme le point, la virgule, le point d'exclamation ou encore la majuscule en début d'énoncé ne sont pas figurées. La segmentation est faite soit sur une unité intuitive de type « groupe de souffle » repérée par le transcripteur humain, soit sur un « tour de parole », défini uniquement par le changement de locuteur.

[21] ELILAP 1980-83 puis LANCOM 1993-2001, voir (Mertens 2002).
[22] Le corpus ESLO est accessible à partir du portail : http://eslo.huma-num.fr/

3.3. Annotation syntaxique d'ESLO

L'annotation syntaxique d'ESLO est une première étape dans son traitement automatique. L'objectif est de permettre aux linguistes (et à d'autres chercheurs des SHS, de mathématiques, d'informatique…) d'effectuer les recherches en utilisant les critères syntaxiques.

3.3.1. Résumé des travaux

L'annotation syntaxique est une étape indispensable dans le traitement automatique du corpus. Mes travaux sur ce type d'annotation ont débuté en 2009, en un temps où aucun outil libre adapté à l'oral n'était disponible. Une réflexion sur la méthodologie à adopter et les jeux d'étiquettes propres à l'oral était indispensable.

Plusieurs possibilités sont envisageables pour l'annotation :

- enlever les disfluences, une technique utilisée par (Valli et Véronis 1999) employée souvent dans le traitement du langage Web

- créer des règles formelles qui prennent en compte les disfluences (Dister 2007, Blanc *et al.* 2008)

- développer un étiqueteur spécifique (Mertens 2002).

Pour respecter au plus près la nature orale des données, je me suis intéressée aux techniques d'apprentissage automatique. Les travaux de l'étiquetage d'ESLO ont été effectués en collaboration avec Isabelle Tellier qui m'a permis de découvrir ce domaine, en particulier les CRF. Il s'agit d'un véritable partenariat dans toutes les tâches entre des compétences linguistiques et informatiques, de la constitution du corpus jusqu'aux tests d'apprentissage et à l'analyse des résultats. Une série de jeu d'étiquettes riches en information linguistique et tenant compte de la nature du corpus traité a été proposée. L'étiqueteur morpho-syntaxique et le chunker ont été appris à partir d'un extrait d'ESLO annoté avec ce jeu d'étiquettes. Le travail effectué a été novateur à l'époque car c'est la première fois qu'un annotateur syntaxique du français a été développé spécialement pour l'oral en utilisant la technique de l'apprentissage automatique. Certaines étiquettes propres à l'oral ont été proposées et la décomposition des étiquettes morpho-syntaxiques par niveaux a donné aussi de bons résultats dans l'adaptabilité des résultats de l'annotation.

La série des travaux sur l'étiquetage morphosyntaxique et le chunking d'ESLO est décrit dans (Eshkol *et al.* 2010, 2012, Tellier *et al.* 2010, 2013, 2014).

Je présenterai dans la partie qui suit la démarche et les réflexions sur les jeux d'étiquettes choisies, les principes et les compromis qui sont souvent élidés ou seulement mentionnées en TAL. Pourtant, ce travail préalable à l'annotation du corpus constitue, au même titre que les résultats quantitatifs communiqués, un apport non négligeable dans le traitement des corpus.

3.3.2. Etiquetage morphosyntaxique

3.3.2.1. Définition et état de l'art

L'étiquetage morphosyntaxique d'un texte est une étape fondamentale de son analyse, et un préliminaire à tout traitement de plus haut niveau. L'objectif de l'étiquetage est d'attribuer à chacun des mots d'un corpus une étiquette qui récapitule ses informations morphosyntaxiques. Le processus d'étiquetage peut accompagner la lemmatisation dont l'objectif est de ramener l'occurrence d'un mot donné à sa forme de base ou « lemme ».

L'étiquetage morphosyntaxique permet d'envisager des recherches non plus sur des formes particulières telles qu'elles se rencontrent dans les textes (chaînes de caractères) mais aussi sur des lemmes (formes canoniques) ou encore sur des catégories syntaxiques. Un corpus du

français parlé annoté avec des informations morphosyntaxiques librement disponible est utile, non seulement pour les logiciels d'annotation en morphosyntaxe, mais également pour améliorer les systèmes de transcription automatique (Huet *et al.* 2006) entre autres.

Il y a des étiqueteurs morphosyntaxiques gratuits ou payants. Le plus connu est TreeTagger[23] (Schmid 1994), un étiqueteur probabiliste qui permet d'annoter un texte avec des informations sur les catégories syntaxiques (POS – *Part Of Speech*) et des informations de lemmatisation. Il n'est pas dédié à une langue particulière et se compose d'un programme principal et de fichiers de paramètres vernaculaires. Parmi les étiqueteurs du français, MEltfr (Denis et Sagot 2009, 2010) utilise d'une part des modèles probabilistes, à savoir des modèles markoviens à maximisation d'entropie et, d'autre part, exploite le lexique Lefff (Sagot 2010). Il existe actuellement une version de MElt pour l'oral du français (fr-perceo), entrainé avec le corpus TCOF-POS (Benzitoun *et al*. 2012).

Ces dernières années, l'étiquetage morphosyntaxique de l'écrit a atteint d'excellents niveaux de performance grâce à l'utilisation de modèles probabilistes et au couplage de ces modèles avec des lexiques externes. Le problème restait entier pour les corpus oraux.

3.3.2.2. Difficultés

3.3.2.2.1. Difficultés « classiques »

Un processus d'étiquetage automatique se trouve confronté aux difficultés suivantes :
- l'ambigüité des mots polycatégoriels où un étiqueteur doit attribuer la bonne étiquette dans un contexte donné

*vous êtes pour ou **contre***
contre [contrer VINDP3S] (à la place de *contre [contre PREP]*)[24]

- des mots non reconnus par des logiciels : mots erronés ou mal orthographiés (*Maing sur Loire* à la place de *Meung sur Loire, traize* à la place de *treize*), des noms propres, des néologismes, des mots étrangers, des abréviations etc.

*les différences qu'il y a entre les lycées […] et les **CES***
CES [ce DETDEM] (à la place de *CES [CES NPPIG]*)

- la segmentation lexicale. La difficulté majeure au cours de ce traitement concerne les mots composés ou les locutions formant les unités lexicales complexes non « segmentables » et contenant un certain degré de non compositionalité lexicale, syntaxique, sémantique et/ou pragmatique. Elles regroupent les expressions figées, les collocations, les entités nommées, les verbes à particule, les constructions à verbe support, les termes etc.

en [en PREP]

effet [effet NCMS]

Dans cet exemple, la locution adverbiale *en effet* n'est pas reconnue : elle se trouve segmentée en deux mots : la préposition (PREP) *en* et le nom commun masculin singulier (NCMS) *effet*.
Dans l'exemple suivant, l'expression *il y a* est segmentée en trois unités distinctes :

il [il PPER3S]

y [y ADV]

a [avoir VINDP3S]

[23] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[24] Les exemples cités ici sont des résultats de l'étiquetage d'ESLO par le logiciel Cordial et les étiquettes de correction proposées ici sont des étiquettes existant dans Cordial.

Les critères linguistiques pour déterminer si une combinaison de mots est une expression figée sont fondés sur des tests syntaxiques et sémantiques décrits dans (Gross 1982, 1996). Ils mettent en jeu : des variations lexicales (*casser / *rompre sa pipe*), des insertions (**casser très fort sa pipe*), des transformations (**sa pipe a été cassée par Max*).

L'identification de ces unités est souvent complexe car elles sont extrêmement hétérogènes en fonction de la variabilité de leur degré de figement. Elles sont difficilement prédictibles automatiquement. Des dictionnaires des mots composés existent comme le DELA (Courtois 2009, Courtois *et al.* 1997) mais ils ne sont pas exhaustifs car ils ne peuvent pas contenir toutes les expressions et locutions du français.

3.1.1.1.1. Difficultés propres au corpus oral transcrit

Pour étiqueter l'oral, d'autres difficultés se retrouvent :

- les disfluences

a). amorce
Dans les conventions d'ESLO, la séquence amorcée est notée par un tiret. L'étiqueteur de l'écrit va l'identifier comme mot composé :

on fait une ou deux ***réclam-*** *réclamations*

réclam- réclamations réclamréclamations NCMIN

à la place de

réclam- reclam- NCI

réclamation réclamation NCFS

Dans l'exemple ci-dessus, la séquence amorcée *réclam-* est étiquetée par le logiciel Cordial ensemble avec la forme qui la suit *réclamation* comme le nom commun invariable en nombre (*NCMIN*), alors qu'il s'agit de deux formes (forme amorcée *réclam-* et forme finie *réclamation*) qui doivent être étiquetées respectivement comme un mot inconnu correspondant dans les étiquettes de Cordial à un nom commun invariable (*NCI*) et un nom commun féminin singulier (*NCFS*).

b). répétition

Le problème de l'étiquetage morphosyntaxique se pose aussi avec les cas de répétition qui sont nombreux à l'oral, surtout si les formes répétées sont polycatégorielles. Observons l'énoncé suivant :

je crois que ***le le*** *les saisons*
le le PPER3S
le le DETDMS

Dans cet exemple, le logiciel annote la première forme *le* comme le pronom personnel à la 3^{e} personne singulier (*PPER3S*) et la deuxième comme le déterminant défini masculin singulier (*DETDMS*). En contexte, on conjecture qu'il s'agit de la répétition du déterminant avant correction.

Dans un autre exemple de répétition :

j'ai été été au cinéma

été VPARPMS

été NCMS

le premier *été* est étiqueté comme le participe passé du verbe *être* au masculin singulier (*VPARPMS*) et le deuxième comme le nom commun au masculin singulier (*NCMS*).

- l'absence de ponctuation dans les fichiers de transcription peut augmenter l'ambiguïté des unités et poser des problèmes au niveau de la segmentation ;

- la présence importante de marqueurs discursifs et d'interjections comme *hein, bon, bien, quoi, comment dire* etc. accroît le nombre de mots inconnus du logiciel et augmente l'ambiguïté comme dans les exemples :

alors ben écoutez madame

ben ben NCMIN

car l'interjection *ben* ne fait pas partie des dictionnaires de Cordial qui est un étiqueteur de l'écrit : il se trouve étiqueté comme nom commun invariant en nombre (*NCMIN*) ; dans :

j'ai quand même des attaches euh ben de la campagne qui est proche quoi

quoi quoi PRI

où *quoi* n'est plus un pronom relatif invariable mais un marqueur discursif. Notons que ces mots, dans leurs emplois en tant que marqueurs discursifs, peuvent être supprimés sans que le sens soit modifié ou remplacé par une interjection. Selon (Dister 2007), « Toute forme peut potentiellement devenir une interjection. On assiste alors à une recatégorisation grammaticale […], le phénomène par lequel un mot ayant une classe grammaticale dans le lexique peut, en discours, changer de classe ». (p. 350)

L'étiqueteur développé avec Isabelle Tellier tient compte de ces spécificités. Le travail sur l'étiquetage morphosyntaxique d'ESLO a été réalisé en deux étapes. Les premières expériences ont été effectuées entre 2009-2010 et décrites dans (Eshkol *et al.* 2010, Tellier *et al.* 2010). J'ai repris ce travail après avec quelques modifications dans le jeu d'étiquettes morphosyntaxiques en 2011-2012 (Eshkol-Taravella *et al.* 2012).

Je reviens à présent sur la méthodologie choisie et sur le jeu d'étiquettes élaboré afin de montrer les particularités et l'apport de mon travail.

3.3.2.3. Méthodologie choisie

La méthodologie générale suit trois étapes :

- sur la base d'un étiqueteur de l'écrit, définir un jeu d'étiquettes répondant au cahier des charges.
- constituer un corpus de référence pour ce nouvel étiquetage.
- entraîner avec ce corpus étiqueté un système d'apprentissage automatique en utilisant les CRF (Conditional Random Fields ou Champs Markoviens Conditionnels).

Le corpus d'entraînement et d'apprentissage doit être « parfait », ce qui implique un travail manuel considérable. Pour amoindrir le coût d'une annotation intégralement manuelle, on a suivi la démarche proposée par (Marcus *et al.* 1993) en procédant à une correction manuelle de corpus pré-annotés automatiquement.

On utilise l'étiqueteur de l'écrit Cordial[25] (Correcteur d'Imprécisions et Analyseur LexicoSyntaxique) (Laurent *et al.* 2009a,b) développé par l'entreprise Synapse. Ce logiciel est également utilisé pour segmenter le corpus et établir le premier ensemble d'étiquettes. Il a été

[25] http://www.synapse-fr.com/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm

choisi pour sa fiabilité et pour sa large palette d'étiquettes, riches d'informations linguistiques. Cordial utilise environ 200 étiquettes indiquant les différentes informations morphologiques comme le genre, le nombre ou l'invariabilité pour les noms et les adjectifs ; la distinction en mode, en temps et en personne pour les verbes ; et même la présence du *h aspiré* en début de mot (Annexe 1).

5 entretiens transcrits (fichiers XML *Transcriber*) convertis en fichiers texte ont été étiquetés d'abord par Cordial (un extrait de la sortie de l'étiquetage de Cordial est montré dans (Annexe 2), puis traités à l'aide de scripts et finalement corrigés manuellement[26] en ajoutant des modifications liées à l'oral (Annexe 5). Le corpus ainsi constitué[27] (18424 mots et 1723 énoncés) a servi de corpus d'apprentissage (Annexe 4).

3.3.2.4. Jeu et format d'étiquettes

3.3.2.4.1. Structure hiérarchique des étiquettes

Les étiquettes morphosyntaxiques portent souvent des informations de nature différente. Au minimum, elles indiquent la partie du discours (POS), i.e. la catégorie syntaxique d'un mot.

Elles peuvent être plus détaillées et inclure d'autres informations de nature :

- morphologique : le genre, le nombre, le temps, le mode etc. ou l'invariabilité pour les formes fléchies ;
- syntaxique : la fonction du mot dans la phrase et les liens qu'il entretient avec d'autres éléments, comme la mention de coordination et subordination pour les conjonctions ;
- sémantique : le caractère possessif, démonstratif, défini, indéfini ou interrogatif pour le déterminant, par exemple.

Pour rendre compte de ces différentes informations, j'ai proposé de structurer les étiquettes sur trois niveaux appelés respectivement L0 (niveau des étiquettes POS), L1 (niveau des variantes morphologiques) et L2 (niveau syntaxico-sémantique), comme dans les exemples ci-dessous :

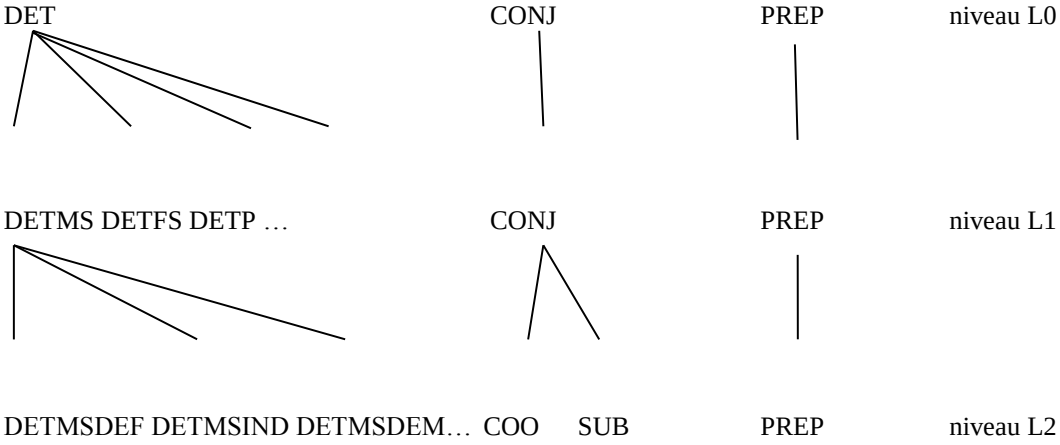

Figure 1*: Structuration hiérarchique de quelques étiquettes*

Comme l'illustre la Figure 1, certaines étiquettes :

[26] La correction manuelle d'un fichier étiqueté par Cordial Analyseur a permis d'établir approximativement le taux d'erreur du logiciel à 4%.

[27] Le corpus de référence a été constitué durant le stage de 3 mois d'étudiants linguistes.

- restent les mêmes sur les trois niveaux pour les catégories non fléchies (les adverbes, les présentateurs, les prépositions etc.) ;
- ne changent qu'au deuxième niveau comme pour les noms, les adjectifs, les verbes, ou qu'au troisième comme pour les conjonctions ;
- varient à chaque niveau en s'enrichissant à chaque fois de nouvelles informations comme pour les pronoms et les déterminants.

La sortie de l'étiquetage se présente en quatre colonnes :

oui	*ADV*	*ADV*	*ADV*
en_effet	*ADV*	*ADV*	*ADV*
on	*P*	*P3I*	*P3IPER*
peut	*V*	*V3SINDP*	*V3SINDP*
commencer	*V*	*VINF*	*VINF*

La première colonne correspond à l'unité lexicale suivie de trois niveaux d'étiquettes.

Cette structure en trois niveaux présente de nombreux avantages. Elle autorise d'abord une certaine souplesse, suivant la nature et la qualité de l'information attendue : le premier niveau est plus simple à étiqueter et donc plus fiable, le troisième niveau inclut des informations linguistiques plus riches mais entraîne potentiellement plus d'erreurs d'étiquetage. On peut ainsi faire des requêtes plus ou moins précises, localisées à un certain niveau. Enfin, on peut varier les expériences d'apprentissage en croisant les différents niveaux.

3.3.2.4.2. Jeu d'étiquettes

Le jeu d'étiquettes comprend 14 étiquettes au premier niveau où n'est indiquée que la catégorie syntaxique (*N, ADJ, DET, P, V, ADV, PREP, CONJ, PRES, MI, INT, MD, CH, UEUPH, PCT*) et 173 étiquettes au dernier niveau où les informations de nature morphologique, syntaxique et/ou sémantiques sont ajoutées (l'Annexe 6). J'ai essayé de tenir compte d'une part, de la nature des données à annoter, et d'autre part, de la tâche finale consistant en un étiquetage automatique par apprentissage. Le compromis trouvé n'est sûrement pas parfait. Nombre de problèmes « classiques » liés à l'ambiguïté ou au nombre important d'étiquettes finales restent à résoudre.

a). Adaptation des étiquettes de Cordial à l'oral

Les 200 étiquettes de Cordial ont été analysées.

Le travail de réflexion sur le jeu d'étiquettes a été mené avec les étudiants du Master dans le cadre des cours « Constitution de corpus », « Enrichissement de corpus ». Il a été poursuivi dans le cadre du stage. Cette étape, primordiale dans le processus de l'annotation, a permis aux étudiants de poser de vraies questions sur la nature des données à annoter, sur les choix à effectuer et sur les décisions à prendre dans les cas où l'interprétation est difficile.

Afin de mieux adapter l'étiquetage à nos besoins, un certain nombre de modifications a été apporté. Ces modifications ont été décidées d'abord en fonction de la nature du corpus, du processus d'homogénéisation des étiquettes et en tenant compte d'une facilitation du processus d'apprentissage automatique. Ces modifications ont été réduites à quatre processus :

- introduction de nouvelles étiquettes : *MI* (mot inconnu) pour les unités non reconnues par l'étiqueteur (les troncations, abréviations etc.) ; *PRES* (présentateur) pour les tournures comme *il y a, c'est, voilà* très présentes à l'oral ;
- simplification des étiquettes de Cordial : la gamme d'étiquettes concernant les invariances de l'adjectif ou du nom (masculin invariant en nombre, féminin invariant en nombre, singulier invariant en genre, pluriel invariant en genre, invariant en nombre et en genre) a été réduite à une seule étiquette (invariable).
- suppression de certaines étiquettes de Cordial : l'étiquette concernant le trait /*h aspiré*/ considérée comme non pertinente ;
- homogénéisation de certaines étiquettes de Cordial : les indications sur le genre et le nombre ont été ajoutées aux déterminants démonstratifs et possessifs par souci de cohérence avec d'autres types de déterminants définis ou indéfinis.

Pour l'étiquetage du corpus ESLO, les 200 étiquettes de Cordial Analyseur ont été ramenées à 114 (Annexe 3).

Plusieurs difficultés à noter. Les étiquettes de Cordial marquent l'invariabilité des mots en genre et en nombre. Le choix a été fait de ne pas préciser cette différence qui n'est pas très pertinente pour l'oral, mais de garder ce trait pour les unités dont la forme ne change pas suivant la conjugaison. Ainsi, le nom *fils* ou *temps* seront étiquetés comme NCI (nom commun invariable), alors qu'un nom comme *madame* recevra une étiquette NCFS (nom commun féminin singulier). Rappelons que les fichiers de transcription n'ont pas été ponctués. Le seul signe typographique conservé est le point d'interrogation. Sans distinguer entre les ponctuations forte et faible faite par Cordial et qui sont pertinentes pour l'écrit, nous avons gardé une étiquette PCT (ponctuation). L'Annexe 5 montre sous forme de tableau l'extrait étiqueté par Cordial tel qu'il a été modifié avec les nouvelles étiquettes.

b). Principes respectés

Principe 1 : tenir compte des spécificités de l'oral

L'objectif principal de la redéfinition des étiquettes consistait dans la volonté de respecter des données orales. Les disfluences de l'oral devaient être traitées le plus finement possible d'où l'introduction de nouvelles étiquettes ou le réaménagement des anciennes.

- Marqueurs discursifs

Les marqueurs discursifs sont des unités lexicales *bon, bien, quoi, comment dire* etc. apparaissant avec une fréquence élevée dans les corpus oraux. Qu'on les désigne comme des *phatiques* et des *régulateurs* (Cosnier 1988, De Gaulmyn 1987), des *particules* (Fernandez 1994), des *marqueurs discursifs* (Roulet *et al.* 1985, Chanet 2001, 2004) ou des *inserts* (Biber *et al.* 1999), ces formes figées ou invariables peuvent constituer des énoncés à elles seules ou s'actualiser en différentes places d'un énoncé sans intégrer sa structure, (c'est-à-dire sans entrer en relation syntaxique avec un autre élément). Elles peuvent donc être supprimées ou modifiées sans que le sens de l'énoncé soit modifié. Leurs fonctions sont très hétérogènes. Elles permettent de temporiser le discours, d'exprimer une marque d'hésitation ou elles peuvent avoir une valeur argumentative. Pour (Chanet 2001), les marqueurs de discours « donneraient des instructions sur la manière dont les interactants peuvent co-construire des représentations, les modifier, et les ajuster les unes aux autres. » Selon la terminologie de l'école de Genève (Roulet *et al.* 1985), les marqueurs de discours peuvent être (i) consécutifs (*alors, donc*), (ii) contre-argumentatifs (*mais*) ou (iii) ré-évaluatifs (*enfin*). Une liste exhaustive est

difficilement réalisable puisque ces mots dépendent de la façon dont ils sont employés et n'importe quelle unité lexicale pourrait changer de statut à l'oral pour être employée en tant que marqueur discursifs.

Pour tenir compte de la variabilité des cas des marqueurs discursifs, j'ai distingué trois catégories :

- marqueurs discursifs propres

*ça a pu arriver mais je me souviens plus **vous savez***
vous_savez vous_savez MD MD MD

- interjections

Ils permettent au locuteur d'exprimer un sentiment ou d'imiter phonétiquement une chose et sont constitués souvent d'un seul phonème vocalique. Il s'agit de mots invariables dont la liste est ici restreinte : *ah, oh, hein, eh, hé, mh, ben* etc.

*un an et demi **ah** peut-être deux ans*
ah ah MD MD MDINT

- marqueur discursif *euh*

(Dister 2007) précise que le *euh* dit d'hésitation constitue d'une part une pause dans le flux de la parole et d'autre part marque une hésitation de la part du locuteur. Il reste au centre de nombreuses études linguistiques (Candéa 2000, Campione et Véronis 2004, Grosjean et Deschamps 1972, 1973, 1975, Morel et Danon-Boileau 1998) ce qui m'a conduite à le distinguer des autres marqueurs.

*les problèmes **euh** littéraires*
euh euh MD MD MDEUH

- Présentateurs

Les présentateurs peuvent faire partie de ce que (Grevisse et Goosse 1993) ont appelé les *introducteurs*. Il s'agit selon les auteurs de mots invariables utilisés pour « introduire un mot, un syntagme ou une phrase ». Cette définition ne s'applique pas intégralement à ce que nous désignons comme *présentateur* car ceux-ci peuvent effectivement être invariables (*voici, voilà, quant à, soit, par exemple, t'as, à savoir, c'est pourquoi, n'est-ce pas etc.*) ou se conjuguer (*c'est, il y a etc.*) :

***voilà** je veux que les Français sachent le français*
voilà voilà PRES PRES PRES

- Mots inconnus

Les mots inconnus dans notre étiquetage sont de deux types :

- des amorces qui sont « la marque d'un achoppement dans l'énonciation, et des recherches qui veulent rendre compte de l'élaboration des énoncés et des modes de productions propres à

l'oral doivent les intégrer dans leurs transcriptions » (Dister 2007). Il s'agit d'un « phénomène langagier qui consiste en « une interruption de morphèmes en cours d'énonciation » (Pallaud 2002:79) Les amorces sont reconnaissables dans la transcription d'ESLO par un tiret.

i- *ils parlent comme nous*
i- *i-* *MI* *MI* *MI*

les ***magi-*** *les magistrats*
magi- *magi-* *MI* *MI* *MI*

- des erreurs de transcriptions :

des avocats qui ***parlement*** *très bien*
parlement *parlement* *MI* *MI* *MI*

faut repasser faut ***raccommoder***
raccommoder *raccommoder* *MI* *MI* *MI*

Principe 2 : simplifier les étiquettes en les rendant moins nombreuses

Le jeu d'étiquettes doit trouver un bon compromis entre d'une part, des étiquettes riches qui reflètent bien les phénomènes linguistiques qu'on veut étiqueter et, d'autre part, des étiquettes en nombre limité car il faut tenir compte du processus automatique de l'apprentissage.

Le nombre d'étiquettes pour l'annotation des catégories syntaxiques est assez stable, ce qui n'est pas le cas pour les informations morphologiques (*niveau L2*) ou sémantico-syntaxique (*niveau L3*). C'est surtout à ces niveaux qu'on observe des différences dans l'étiquetage morphosyntaxique.

Pour diminuer le nombre d'étiquettes, j'ai supprimé certains traits linguistiques repérables autrement ou discutables du point de vue théorique.

- Genre et nombre des noms propres

Dans la catégorie des noms, deux classes ont été distinguées : noms communs et noms propres. Cette distinction traditionnelle est présente dans tous les discours de référence. Les conventions de transcription d'ESLO interdisent l'utilisation de signes typographiques sauf pour les noms propres marqués par la majuscule, ce qui les rend facilement repérables. Le problème concerne les informations de nature morphologique. Comment identifier le genre et le nombre des noms propres qui s'utilisent sans article *Orléans, Paris.* (Riegel 1994:175) constate ce fait sans pour autant expliquer la différence « Les noms propres s'écrivent avec une majuscule, n'ont pas de déterminant (*Pierre, Paris*) ou bien se construisent avec un déterminant contraint, l'article défini (*Le Rhin, les Vosges*). » Remarquons de plus que cet article peut faire partie du nom propre *Le Loiret* (dans ce cas il est en majuscule selon les conventions de transcription d'ESLO) mais pas *la Loire*.

Suite à cette absence d'homogénéité des emplois et pour éviter les erreurs d'étiquetage, j'ai supprimé les traits concernant la variabilité des noms propres. Ils ont une seule étiquette (NP) qui ne varie pas suivant les niveaux.

Citroën Citroën NP NP NP

➢ Déterminant partitif

Je distingue cinq types de déterminants : définis, indéfinis, possessifs, démonstratifs et interrogatifs. La question s'est posée pour l'article partitif : faut-il l'étiqueter séparément ? Dans ce cas, il faudrait distinguer entre l'article partitif et contracté :

*votre mari a **de la** famille / tout va dépendre **de la** famille de mon mari*

Dans le premier cas, *de la* est l'article partitif, dans le deuxième, il s'agit de la préposition suivie de l'article défini.

(Galmiche 1986) indique que « l'article dit partitif (du, de la, de l') se trouve davantage apparenté à l'article indéfini pluriel (des) qu'à l'article indéfini singulier (un). » (Riegel 1994), à son tour, note que l'article partitif *du, de la* ou *de l'* s'emploie devant « le singulier des noms de masse (du plâtre, de la farine) et des noms dits « abstraits » qui ne renvoient pas à des entités comptables (du courage, de la lâcheté) ». La forme pluriel *des* est utilisée « avec des termes massifs essentiellement pluriels […] des décombres, des épinards etc.) » De nombreux travaux traitent la notion et l'emploi du partitif (Anscombre 1996, Attal 1976, Englebert 1996, Galmiche 1986, Kupferman 1994, 1996, 1998, 1999, 2001). Les études varient en fonction des écoles théoriques et les critères caractéristiques restent flous et non homogènes, c'est pourquoi les formes *du, de la, de l', des* dans leur emploi « partitif », c'est-à-dire lorsqu'elles indiquent « une partie de … », sont incluses dans la classe des déterminants indéfinis. La définition des indéfinis donnée dans le *Bon Usage* de (Grevisse et Goosse 1993) va dans le même sens. Les auteurs notent que les indéfinis indiquent « soit une quantité non chiffrée, soit une identification imprécise ou même un refus d'identification ».

On trouvera ainsi parmi les déterminants indéfinis :

- les articles indéfinis (*un, une, des*),
- les articles partitifs (*du, de la, des, de l'*),
- les unités lexicales appelés par certains grammairiens « adjectifs indéfinis » (*aucun, certain, quelque etc.*),
- des déterminants quantitatifs (*beaucoup de, pas mal de etc.*).

*est-ce que vous faites **un** brouillon*
un un DET DETMS DETMSIND

*on prend n'importe quoi oui **n'importe quel** papier*
n'importe_quel n'importe_quel DET DETMS DETMSIND

*votre mari a **de la** famille*
de la de la DET DETFS DETFSIND

*y a **trop de** circulation*
trop_de trop_de DET DETI DETIIND

➢ Adverbes

La classe des adverbes est très hétérogène. (Riegel 1994:375) définit les adverbes par exclusion « les termes invariables qui ne sont ni des prépositions ni des conjonctions ni des interjections ». L'auteur cite trois critères généralement proposés pour les adverbes : l'invariabilité, le caractère facultatif et la dépendance par rapport à un autre élément de l'énoncé. Je ne vais pas discuter de ces critères dès lors que les deux derniers (la facultativité / la dépendance) se contredisent.

La classe d'adverbes est une classe très hétérogène qui regroupe différentes sortes de mots difficiles à définir. La question se pose également pour leur classement. Faut-il distinguer entre les adverbes interrogatifs, de temps, de manière etc.?

J'ai choisi de ne distinguer qu'un seul sous-groupe au sein de cette classe, les adverbes de négation. D'une part, ces adverbes peuvent être énumérés car ils contiennent un nombre fini d'unités lexicales et d'autre part, une étiquette spécifique (*ADVNEG*) permettrait de repérer plus facilement les énoncés négatifs (ceux où la négation est exprimée à travers les adverbes).

*Je **ne** pense **pas***
ne ne ADV ADV ADVNEGDISC1
pense penser V V1SINDP V1SINDP
pas pas ADV ADV ADVNEGDISC2

- Adjectifs invariables

La classe des adjectifs pose le problème du trait « invariable » présent dans les étiquettes de Cordial. L'adjectif s'accorde en genre et en nombre avec le nom auquel il se rattache. Cependant, il existe les adjectifs dont la forme ne change pas selon le genre comme *brave*, *rouge* etc. (Riegel 1994) les appelle « adjectifs à forme unique à l'oral et à l'écrit ». Ce problème ne se pose pas par contre pour le nombre où le *s* est toujours ajouté en pluriel.

Il a été décidé d'attribuer à l'adjectif une étiquette indiquant son genre qui se déduit, dans le cas des adjectifs non variables, du genre du nom qu'il accompagne.

*L'école primaire est **indispensable** (ADJFS) […] c'est **indispensable** (ADJMS). »*

Le premier emploi de l'adjectif *indispensable* s'accorde avec *l'école primaire*, donc il reçoit le marque de féminin (*ADJFS*) ; alors que le deuxième est utilisé avec le présentatif *c'est*, où le pronom *ce* est par défaut masculin singulier.

3.3.2.4.3. Particularités de l'étiquetage proposé

- Nouvelles étiquettes

Comme il a été mentionné dans la section précédente (3.3.2.4.2), quelques nouvelles étiquettes liées à la nature des données traitées comme *PRES* (présentateur) et *MD* (marqueur discursif) avec deux sous-groupes *MDEUH* (*euh* d'hésitation) et *MDINT* (interjection) ou encore *MI* (mot inconnu) ont été ajoutées.

- Discontinuité

Les étiquettes proposées tiennent compte de la discontinuité de certaines structures en indiquant leurs frontières (*DISC1* et *DISC2*). En premier lieu, il s'agit des constructions négatives :

Je ***ne*** *pense* ***pas***
ne ne ADV ADV ADVNEGDISC1
[...]
pas pas ADV ADV ADVNEGDISC2

où la négation est marquée par les deux adverbes séparés *ne* et *pas* disjoints par l'insertion médiane du verbe. Un autre exemple de constituant discontinu est la restriction *ne ... que*, une structure proche de la négation où l'adverbe *ne* n'a pas de valeur négative :

il ***ne*** *vous reste* ***qu****'un enfant*
ne ne ADV ADV ADVDISC1
[...]
qu' que ADV ADV ADVDISC2

Les présentateurs *c'est* ou *il y a* peuvent être décomposés lorsqu'ils sont à la forme négative, par exemple. Dans ce cas, on indique deux structures discontinues en chevauchement :

ce n'est plus *la même vie*
ce c'est PRES PRES PRESDISC1
n' ne ADV ADV ADVNEGDISC1
est c'est PRES PRES PRESDISC2
plus plus ADV ADV ADVNEGDISC2

ou lorsqu'un clitique figure au milieu :

y'en aura *plus c'est fini*
y il_y_a PRES PRES PRESDISC1
en en P P3S P3SPERCOMPL
aura il_y_a PRES PRES PRESDISC2

- Possession

Pour étiqueter les déterminants et pronoms possessifs, on a tenu compte du « possesseur » (celui qui possède) et du « possédé » (l'objet possédé), comme indiqué dans le Tableau 1.

Des exemples d'étiquetage de ces unités sont présentés ci-dessous :

je m'occupe de ***mon*** *intérieur*

mon	*mon*	*DET*	*DETMS1S*	*DETMS1SPOSS*

ses *études primaires*

ses	*son*	*DET*	*DETP3S*	*DETP3SPOSS*

c'est ***le sien***
le_sien le_sien P PMS3S PMS3SPOSS

Possesseur = Qui possède	Possédé=Objet possédé			
	Singulier		Pluriel	
	masculin	féminin	masculin (pour les pronoms)	féminin (pour les pronoms)
1S	*mon/le mien*	*ma/la mienne*	*mes/les miens*	*les miennes*
2S	*ton/le tien*	*ta/la tienne*	*tes/les tiens*	*les tiennes*
3S	*son/le sien*	*sa/la sienne*	*ses/les siens*	*les siennes*
1P	*notre/le notre*		*nos/les nôtres*	
2P	*votre/le votre*		*vos/les vôtres*	
3P	*leur/le leur*		*leurs/les leurs*	

Tableau 1 : Tableau récapitulatif des déterminants et pronoms possessifs

- Multi-mots ou locutions

Le processus d'étiquetage est lié directement à celui de la segmentation qui se complique avec les mots composés, les expressions polylexicales ou les expressions multi-mots (le terme inspiré du terme anglais « multi-word expression ») formant les unités lexicales complexes non « segmentables » et contenant un certain degré de non compositionalité lexicale, syntaxique, sémantique et/ou pragmatique. La gamme des phénomènes est variée, et, malgré un relatif consensus autour de certaines catégories, il n'existe pas de typologie faisant l'unanimité, ni même de délimitation claire de l'ensemble des constructions concernées. Elles regroupent les expressions figées, les collocations, les entités nommées, les verbes à particule, les constructions à verbe support, les termes etc. Elles peuvent être contiguës (*cordon bleu, par rapport à etc.*) ou pas (*donner un avertissement, prendre en compte*), opaque sémantiquement (*tout à fait*) ou pas (*vin blanc*).

Il existe donc une grande variété de phénomènes linguistiques rentrant dans cette catégorie et de nombreux critères d'identification. La reconnaissance automatique des unités multi-mots est, la plupart du temps, réalisée à l'aide de ressources lexicales construites manuellement comme dans le cas des expressions figées (Gross 1997, Silberztein 2000) ou apprises automatiquement comme dans le cas des collocations nominales (Seretan *et al.* 2003). L'identification de telles expressions est une tâche difficile car les unités non décrites dans les ressources sont difficilement identifiables.

La liste des mots composés et des expressions que nous avions à disposition provenait du premier étiquetage automatique effectué et correspondait aux expressions multi-mots considérées comme telles par Cordial. Il fallait donc retravailler cette liste pour supprimer les expressions « segmentables », d'une part, et d'autre part y ajouter de nouveaux items. Plusieurs critères ont été mis en œuvre pour distinguer les multi-mots :

- les mots composés sont des séquences non compositionnelles de mots ;
- ils peuvent être contigus, mais ce n'est pas toujours le cas ;
- les composants de l'expression ont perdu leur sens d'origine ;

- ils peuvent être traduits dans une autre langue par un mot simple (ce critère s'applique souvent aux structures à verbe support ou aux constructions avec les verbes à particules).

Ces critères ne sont pas exhaustifs. Quelques exemples d'expressions multi-mots pour les catégories utilisées sont présentés ci-dessous.

- Noms composés :

  chemin de fer, bureau de poste, point de vue, cul de poule, école primaire, stylo bille, sac à main, femme au foyer, bonne sœur

- Adverbes composés :

  par cœur, tambour battant, à tous les coups, à bout de bras, tout le temps, quand même, un peu, un petit peu, pas du tout

- Conjonctions composés :

  dès que, même si, parce que

- Pronoms composés :

  tout le monde, tout à l'heure
  qui est-ce que, qui est-ce qui, qu'est-ce que, qu'est-ce qui, qu'est-ce qu'

- Verbes composés

  V+N : *(se) rendre compte, avoir besoin, avoir mal, faire attention*
  V+ADJ : *avoir beau*
  V+PREP+N/GN : *être à cheval sur, être de passage, être en train de*
  V+VINF : *entendre dire*

- Marqueurs discursifs :

  mon dieu

- PRES (présentateur)

  c'est, il y a

Il n'y a pas toujours de consensus entre les chercheurs concernant les expressions multi-mots. Ainsi, le fait d'être séparé par un autre mot est une caractéristique éliminatoire dans le cas de l'annotation du corpus TCOF présenté dans (Benzitoun *et al.* 2012) où « toute séquence dans laquelle il est possible d'insérer un élément est découpée en plusieurs tokens, afin d'exclure les unités discontinues. Ainsi, *un peu* est découpé en deux tokens (car on peut trouver *un tout petit peu*) ». Dans ce cas, notre étiquetage marque le lien entre les deux unités par le trait *DISC* (discontinuité) :

*J'**ai besoin** de ça*

ai_besoin	*avoir_besoin*	*V*	*V1SINDP*	*V1SINDP*

*J'**ai** vraiment **besoin***

ai	*avoir_besoin*	*V*	*V1SINDP*	*V1SINDPDISC1*
vraiment	*vraiment*	*ADV*	*ADV*	*ADV*
besoin	*avoir_besoin*	*V*	*V1SINDP*	*V1SINDPDISC2*

- Fonction syntaxique

La volonté de rédiger des étiquettes informatives et de faciliter le traitement syntaxique postérieur, ont conduit à ajouter aux pronoms des indications sur leurs fonctions syntaxiques. Cet ajout assure la désambiguïsation comme dans les exemples ci-dessous où *nous* a plusieurs fonctions :

***Nous** regardons (sujet)*
*Nous **nous** regardons (complément direct, réfléchi)*
*Ils **nous** regardent (complément direct)*
*Ils parlent avec **nous** (complément indirect)*

Ainsi, la distinction entre le sujet et le complément toniques a été introduite :

*et **vous** y comptez rester ?*
vous vous P P2P P2PPERSUJ

*je pourrais **vous** demander*
vous vous P P2P P2PPERCOMPL

*je parle avec **vous***
vous vous P P2P P2PPERTON

Les pronoms *en* et *y* sont étiquetés comme pronoms adverbiaux :

*je n'**y** vais pas*
y y P P3S P3SPERADV

- Passif

L'information sur le passif est importante à distinguer pour permettre de retrouver plus facilement cet emploi dans le corpus. Une raison supplémentaire est d'éviter une ambiguïté possible concernant les énoncés au passé composé avec l'auxiliaire *être* et le passif :

Je suis prise / je suis partie

Sans précision du passif, il faudrait rechercher le verbe auxiliaire *être* suivi d'un ou de deux participes passés ce qui extrait en bruit les énoncés avec le passé composé.

*j'ai été **mutée***
j' je P P1S P1SPERSUJ
ai avoir V V1SINDP V1SINDPAUX
été être V VMSPP VMSPP
mutée muter V VFSPP VFSPPPAS

- Temps composés

Une stratégie différente a été adoptée pour les temps composés. Les étiquettes propres aux temps composés (*passé composé, plus-que-parfait, futur antérieur etc.*) n'ont pas été créées car il est possible de les retrouver en cherchant l'auxiliaire à un temps précis (*présent, imparfait, futur etc.*) suivi d'un participe passé.

*elle **a disparu***
a avoir V V3SINDP V3SINDPAUX
disparu disparaître V VMSPP VMSPP

3.3.2.4.4. Problèmes du jeu d'étiquettes proposé

- Ambiguïté

Le problème classique de l'étiquetage est l'ambiguïté entre les différents emplois d'une même unité lexicale. Ce problème est difficile à résoudre dans le cas d'un étiquetage automatique limité au contexte immédiat et qui ne peut recourir à des connaissances extralinguistiques.

- les pronoms relatifs et interrogatifs

Cette ambiguïté concerne des formes comme *qui, que, quoi, lequel* et ses variantes. On sait qu'à l'écrit le pronom interrogatif est placé souvent en début de phrase :

Lequel de ces plats veux-tu ?

On peut également le retrouver seul, en mot-phrase dans une conversation :

Je possède déjà l'un de ces tableaux. – Lequel ?

Cependant, à l'oral, où la notion de phrase est mise en question et où les marques typographiques (comme le point ou la majuscule en début de phrase) sont absentes, les deux emplois sont plus difficiles à distinguer. Ces pronoms interrogatifs peuvent être employés n'importe où dans l'énoncé :

c'est dans une boîte d'accord mais dans laquelle ?

dans laquelle c'est ?

- certains pronoms personnels

Les pronoms tels que *nous, vous* peuvent remplir différentes fonctions syntaxiques dans l'énoncé : sujet, complément ou apostrophe et apposition, ce qui pose évidemment un problème lors de l'étiquetage.

Lorsqu'ils sont utilisés après une préposition, ils sont forcément toniques :

est-ce que vous aviez un dictionnaire chez vous ?

mais malheureusement le contexte ne peut pas toujours être aussi tranché.

- les marqueurs discursifs

A l'oral, certains verbes cognitifs comme *voir, tenir, comprendre etc.* peuvent être employés dans leurs forme impérative non pour exprimer un ordre ou un conseil mais pour attirer l'attention de l'interlocuteur ou simplement reprendre une idée. Dans ce cas, ils sont des marqueurs discursifs. Le test de suppression de ces formes sans que la syntaxe et le sens de l'énoncé soit atteint aide à résoudre ce problème dans l'étiquetage manuel mais il s'avère impraticable en étiquetage automatique.

c'est devenu français ça ***voyez*** *ce sont des mots qui sont passés dans la langue*
voyez voyez MD MD MD

La conjonction de coordination *donc* n'est pas toujours employée pour la coordination. A l'oral, elle peut avoir un autre emploi, celui d'un marqueur discursif :

c'est tout ce que je connais ***donc*** *je m'y plais*
donc donc CONJ CONJ CONJCOO

aujourd'hui c'est plus ça et puis regardez ***donc*** *autrefois on apprenait l'alphabet*
donc donc MD MD MD

On retrouve souvent ce type d'emploi en fin de question :

comment s'appelle-t-il ***donc***
donc donc MD MD MD

L'adjectif *bon* peut aussi changer de catégorie grammaticale et être employé à l'oral comme un marqueur discursif :

on n'était pas toujours ***bon*** *en français*
bon bon ADJ ADJMS ADJMS

oui ***bon*** *alors pour revenir un peu à l'enseignement madame*
bon bon MD MD MD

Cependant, on peut les distinguer grâce au contexte de l'énoncé :

En tant qu'adjectif :

- *bon* détermine le nom qui le suit ou le précède ;
- *bon* suit le verbe attributif (*il est bon*).

En tant que marqueur discursif, il est souvent accompagné d'un adverbe ou d'un marqueur discursif avant ou après *bon* : *oui, ah oui bon, bon alors*.

Un autre cas d'ambiguïté concerne la séquence *bien que* qui peut être analysée soit comme une conjonction de subordination, soit comme deux unités distinctes comme dans l'exemple :

je crois bien que le gouvernement

Il n'y a pas de règles qui permettent de différencier ces deux emplois. Ce n'est que la prosodie, c'est-à-dire l'écoute du fichier sonore, qui pourrait indiquer la différence.

- Répétition

(Henry 2005) distingue :

- les répétitions « faits de langue » où la répétition est due à la syntaxe ;
- les répétitions « faits de parole » qui font partie des disfluences de l'oral.

Ce phénomène concerne souvent les cas de répétition des pronoms personnels. Si dans le premier cas, les deux éléments reçoivent deux étiquettes différentes :

nous nous habillons
nous nous P P1P P1PPERSUJ
nous nous P P1P P1PPERCOMPL

car ils occupent deux fonctions syntaxiques différents, dans le deuxième, il s'agit plutôt d'une disfluence :

ça traduit toute une euh nous nous n'étions pas comme ça

3.3.2.5. Apprentissage

Isabelle Tellier et moi soutenons que l'apprentissage automatique doit exploiter au maximum les connaissances linguistiques disponibles sur le corpus. La méthode d'annotation proposée est guidée par la nature des données, la technique et les expériences de l'apprentissage automatique le sont aussi.

L'apprentissage a utilisé d'abord la structure hiérarchique des étiquettes en testant diverses stratégies de décomposition des étiquettes en sous-étiquettes plus simples. D'autres types de connaissances linguistiques ont été ensuite pris en compte tels que la structure morphologique de l'unité qui décompose les mots en constituants, distinguant une racine et une séquence de lettres finales, souvent porteuses de certaines informations morphologiques : des désinences comme *-ait, -ais, -is, -é, -s*, indiquent, le temps verbal, le genre et le nombre etc., i.e. les morphèmes grammaticaux. En considérant la racine comme la partie commune à toutes les formes d'un mot, on extrait ces séquences finales de la forme de surface pour déterminer la partie morphologique de l'étiquette qui doit être associée au mot. Les expériences d'apprentissage automatique sur le corpus de référence annoté manuellement avec le jeu d'étiquettes établi sont décrites dans (Eshkol *et al.* 2010, Tellier *et al.* 2010, Eshkol *et al.* 2012).

3.3.2.6. Conclusion

Résultats

Les expérimentations ont montré que :

- la méthodologie des techniques d'apprentissage automatique est prometteuse et pertinente dans le cas de corpus non normalisés comme ceux de l'oral[28] ;

- la technique d'apprentissage automatique permet de respecter les données car elle est basée sur le corpus d'exemples annotés manuellement. Elle tient compte des choix faits antérieurement par les linguistes.

- les CRF est un modèle statistique riche qui permet d'exploiter efficacement les différentes informations fournies (le contexte, les informations linguistiques, les ressources extérieures, les règles).

- la structure hiérarchique des étiquettes permet une souplesse d'étiquetage et d'usage et également de l'apprentissage.

- il est possible d'atteindre 89 % d'exactitude (au troisième niveau) à 94 % (au premier niveau) pour un étiqueteur. Ces résultats sont comparables à ceux obtenus sur le corpus TCOF (Benzitoun *et al.* 2012) : 96,9% avec MElt et 94,9% avec TreeTagger. Ils sont pourtant inférieurs aux performances annoncées par les meilleurs étiqueteurs actuels du français écrit, qui obtiennent entre 97 et 98 % d'exactitude (Denis et Sagot 2010, Constant *et al.* 2011) avec un jeu d'étiquettes comparable à notre premier niveau, mais ces étiqueteurs ont été entraînés sur le French TreeBank, soit plus de 300 000 unités lexicales extraites d'articles du Monde, dans une langue normée. L'oral présentant plus d'irrégularités, il est à prévoir que les modèles statistiques entraînés sur ESLO requièrent plus d'exemples encore pour parvenir à des résultats équivalents.

Contraintes

Plusieurs contraintes ont été respectées durant ces expériences :

- l'objectif étant de développer l'étiqueteur de l'oral, nous avons mis de côté deux tâches importantes dans le processus d'annotation morphosyntaxique : la lemmatisation et la segmentation et nous nous sommes basés sur les résultats du logiciel Cordial ;

[28] (Benzitoun et al. 2012) a adopté la même méthodologie pour étiqueter le corpus TCOF par la suite.

- pour faciliter les expériences, les fichiers de transcription sur lesquels nous avons travaillé ont été nettoyés des balises XML indiquant les métadonnées et les informations temporelles. Nous avons réussi par la suite d'étiqueter ces fichiers en gardant les balises XML.

Conséquences

Les travaux sur l'étiquetage d'ESLO ont été commencés en 2009. Ce travail se place du point de vue chronologique après (Dister 2007)[29] et avant celui de TCOF-POS (Benzitoun *et al.* 2012) et DisMo (Christodoulides *et al.* 2014). Le tableau de différents jeux d'étiquettes établies pour quatre étiqueteurs (en incluant ESLO) est présenté dans l'Annexe 7. Cela permet de visualiser les choix théoriques faits par les chercheurs travaillant sur une même tâche.

Ces différents systèmes d'étiquetage morphosyntaxique du français parlé convergent vers :

- un passage à l'apprentissage automatique

Les trois derniers systèmes ont utilisé la méthode par apprentissage qui, à l'heure actuelle, semble donner de meilleurs résultats. Le temps et le coût de constitution du corpus de référence peuvent être réduits par la pré-annotation du corpus au moyen d'un étiqueteur avant d'être modifié à l'aide de scripts.

- la réduction du nombre d'étiquettes

On constate que le nombre d'étiquettes influence la qualité d'étiquetage développée par un apprentissage automatique. Moins d'étiquettes à apprendre facilite la tâche.

Le nombre élevé d'étiquettes accroît la levée d'ambiguïtés. Plus on veut distinguer d'emplois d'une unité lexicale, plus le logiciel a de possibilités, source d'autant de difficultés. La réduction du nombre d'étiquettes passe par la suppression d'informations morphologiques sur le genre, le nombre et la personne. Certaines étiquettes, surtout lorsqu'elles ne s'appliquent qu'à une seule unité, se justifient difficilement.

- la volonté de prendre en considération l'oral

On observe l'affirmation d'un intérêt croissant pour l'oral. Si dans la thèse d'Anne Dister, aucune étiquette (sauf une interjection) ne reflétait l'oral, l'étiquetage appliquant plutôt les étiquettes classiques utilisées à l'écrit, les logiciels développés ultérieurement ajoutent au jeu d'étiquettes certaines propres à l'oral :

5 étiquettes dans ESLO (*PRES, MI, MD, MDEUH, MDINT*);
8 étiquettes dans TCOF-POS (*MLT, TRC, LOC, V :trc, ADJ :trc, NOM :trc, NAM :trc, FNO)*
11 étiquettes dans DisMo (*AMO, CORR-B, CORR-I, REP-B, REP-I, SIL:b, SIL:l, SIL:s, HESI, MD, PARA*)

On constate une présence accrue des phénomènes propres à l'oral dans la définition des étiquettes.

En ce qui concerne ESLO, la prise en compte de toutes les informations morphologiques a augmenté significativement le nombre d'étiquettes ce qui a rendu l'apprentissage automatique difficile et a diminué la qualité des résultats, surtout au troisième niveau. La prise en compte aurait pu être encore plus extensive si l'on avait introduit les étiquettes *FNO* (forme noyau) pour des énoncés comme *oui, non, d'accord*, très présents à l'oral, surtout en interview.

[29] Il s'agit de l'étiqueteur développé par Anne Dister dans le cadre de sa thèse pour le corpus Valibel.

Cependant, l'ajout de l'étiquette *PRES* (présentateur) pour les séquences comme *c'est, il y a, voilà etc.* semble être une bonne décision de notre part alors qu'elle ne figure pas dans les jeux d'étiquettes d'autres logiciels.

Il me semble intéressant d'harmoniser les travaux portant sur le processus d'étiquetage syntaxique de l'oral pour arriver à un consensus collectif sur le jeu final d'étiquettes et sur les phénomènes de l'oral qui doivent y apparaître. Plusieurs initiatives vont dans ce sens : le Consortium des corpus oraux, ORTOLANG etc. Une journée de travail sur l'annotation morphosyntaxique des corpus oraux a été organisée le 13 mars 2014 par les participants au projet PFC. Elle a réuni plusieurs chercheurs qui ont partagé leur expérience du travail en ce domaine.

3.3.3. Analyse syntaxique en chunks

La suite du travail sur l'annotation syntaxique d'ESLO a porté sur la segmentation en chunks. Cette analyse a fait ses preuves sur l'oral (Antoine *et al.* 2003, Blanc *et al.* 2010, Paroubek *et al.* 2007). Selon (Antoine *et al.* 2003), la segmentation en chunks présente de nombreux avantages. Elle « garantit une certaine robustesse tout en autorisant une analyse plus détaillée des énoncés oraux. Contrairement aux approches sélectives, aucun élément n'est en effet ignoré à ce stade. De même, 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. » La tâche de chunking est fondée sur un étiquetage en POS préalable.

Celle que nous avons effectuée a été décidée en fonction de plusieurs objectifs et de plusieurs contraintes. Tout d'abord, nous avons voulu chunker le mieux possible les transcriptions de l'oral. Deux contraintes ont pesé sur le choix méthodologique : la taille restreinte du corpus oral de référence et la volonté d'économiser le coût d'annotation en minimisant l'intervention humaine. Pour respecter ces impératifs, les tests d'apprentissage ont été réalisés sans recours à un nouvel étiqueteur POS.

J'ai montré dans la partie précédente que l'apprentissage automatique supervisé utilisant les CRF est particulièrement performant pour une tâche d'annotation syntaxique en surface mais qu'il dépend fortement du corpus de référence sur lequel il est appris. Nous avons voulu utiliser la même technique de l'apprentissage mais il fallait l'adapter aux nouvelles données. Nous disposions d'un étiqueteur POS et d'un chunker (SEM) appris à partir d'une grande quantité de données écrites annotées (le FTB) et nous souhaitions chunker des données nouvelles très différentes par leur source.

Nous avons procédé en deux étapes :

- application au corpus oral du chunker utilisé sur les données écrites ;
- apprentissage de la tâche de chunking à partir des données orales annotées en chunks adaptés.

Ces travaux sont décrits dans (Tellier *et al.* 2012, 2013, 2014).

3.3.3.1. Chunking : définition et état de l'art

Les dernières décennies, un nouveau système, les *shallow parsers* (analyseurs peu profonds) a été développé pour l'analyse syntaxique. Aussi appelés *chunkers*, l'objectif de ces parseurs est de segmenter l'énoncé en constituants minimaux (chunks) tout en analysant leur structure interne. Il s'agit d'une analyse syntaxique qui se base sur les parties du discours, donc sur un étiquetage morphosyntaxique préalable. Certains analyseurs sont restreints à la simple

identification des *chunks* (les constituants syntaxiques les plus petits dans la phrase), alors que d'autres peuvent indiquer des relations syntaxiques.

Les chunks sont des constituants continus et non-récursifs (Abney 1991). Ils définissent la structure syntaxique superficielle des phrases et, à ce titre, sont moins coûteux et plus faciles à obtenir que la structure en constituants complète. Pour certains textes non normés (transcriptions de l'oral par exemple), ils représentent le degré d'analyse le plus poussé qu'on puisse espérer. Il a en effet été démontré que ces constituants sont le lieu de réalisation privilégié des réparations à l'oral (Blanche-Benveniste 1997:47).

La notion de chunk n'est pas toujours très précisément définie et peut recouvrir plusieurs niveaux de détails possibles, suivant que l'on se concentre sur :

- les groupes nominaux non récursifs à la façon de (Sha et Pereira 2003), incluant les éventuels groupes adjectivaux immédiats, déterminants et adjectifs numériques. Les compléments du nom font partie de chunks distincts de celui du nom qu'ils qualifient.
- l'ensemble de tous les constituants non récursifs possibles : les groupes nominaux, verbaux, prépositionnels, adjectivaux etc.

Ainsi, le même énoncé

la situation de mon fils est difficile

- peut être segmenté en chunks nominaux sans intégrer les unités lexicales qui n'en font pas partie :

(la situation)$_{NP}$ *de (mon fils)*$_{NP}$ *est difficile*

- peut être segmenté en autant de chunks, tous les éléments étant inclus dans l'un des chunks :

(la situation)$_{NP}$ *(de mon fils)*$_{PP}$ *(est)*$_{VN}$ *(difficile)*$_{AP}$

Pour aborder la tâche de chunking comme une tâche d'annotation, il suffit d'associer à chaque mot appartenant à un chunk une étiquette donnant son type (soit NP, soit un type parmi {NP, VN, PP, AP, AdP, VCOOR}) visualisé à l'aide de parenthèses comme dans l'exemple ci-dessus ou accompagné du codage *BIO (Begin/In/Out)* qui permet de délimiter ses frontières.

3.3.3.2. Données et outils de base

3.3.3.2.1. SEM

Plusieurs méthodes et outils ont été proposés pour le chunking :

- une analyse syntaxique superficielle de textes non normés (Blanc *et al.* 2010).
- les systèmes ayant participé aux campagnes d'évaluation Easy et Passage (Paroubek *et al.* 2007) ;
- une plateforme généraliste et multilingue comme Gate[30] ;
- les systèmes appris automatiquement à l'aide d'un CRF (Sha et Pereira 2003).

A ma connaissance, peu de solutions spécifiques et gratuites sont disponibles pour le chunking du français oral.

Yoan Dupont, un étudiant d'Isabelle Tellier, a développé sous sa direction le logiciel SEM (Constant *et al.* 2011, Constant et Tellier 2012, 2013), un étiqueteur en POS fondé sur

[30] http://www.semanticsoftware.info/munpex

l'apprentissage automatique à l'aide des CRF et appris sur le corpus écrit French Tree Bank (Abeillé *et al.* 2003).

Nous avons voulu ajouter une couche supplémentaire d'annotation syntaxique en intégrant la segmentation du texte en chunks. J'ai pu contribuer à la réflexion sur la définition des chunks et tester ce logiciel sur les données orales.

3.3.3.2.2. Sous-corpus oral ESLO

Le corpus de test correspond à un échantillon du corpus ESLO1 (les entretiens) constitué de 852 tokens étiquetés en POS par un étiqueteur décrit dans la section 3.3.2. Deux niveaux de chunking (comprenant les groupes nominaux et l'ensemble des groupes non récursifs) ont été testés et évalués au cours de tests.

3.3.3.3. De l'écrit vers l'oral : jeu d'étiquettes et résultats

3.3.3.3.1. Choix des chunks[31] dans SEM

Nombre et nature des chunks

Le choix des chunks dans SEM a été directement associé aux étiquettes POS de FTB (Crabbé et Candito 2008).

Etiquettes POS	Etiquettes de chunks
NC, NPP	NP
ADJ, ADJWH	AP
DET, DETWH	
ADV	AdP
P	PP
CC, CS	CONJ
PRO, PROREL, PROWH, CLS, CLO, CL	NP
V, VINF, VS, VPR, VPP	VN
I	AdP
UNKNOWN	UNKNOWN

Tableau 2 : Sept chunks de SEM

[31] Le travail sur le guide d'annotation a été mené dans le cadre du stage d'Ilaine Wang que j'ai co-encadré avec Isabelle Tellier (http://www.lattice.cnrs.fr/sites/itellier/guide.html).

Il s'agit de sept chunks et de six grands types de groupes (têtes potentielles) (voir Tableau 2).

Ce choix nécessite quelques remarques :

- pronoms

Les pronoms (sauf les pronoms qui remplacent des groupes prépositionnels *y, en, dont, où*), y compris les pronoms clitiques sujets (*CLS*), objets (*CLO*) et réfléchis (*CLR*), ne sont pas compris dans les groupes verbaux (*VN*) et sont considérés comme faisant partie du chunk nominal (*NP*). Ce choix, discutable du point de vue théorique, a été dicté par l'objectif d'une distinction à réaliser entre les pronoms afin de faciliter l'accès à l'information en cas de coréférence :

(***je***/***B-NP***) (***me***/***B-NP***) (*plais*/*B-VN*)
(***il***/***B-NP***) (***y***/***B-VN*** *a*/*I-VN*)

- adjectifs

Les adjectifs épithètes, quels que soient leur nombre et leur position, font partie d'un chunk nominal gouverné par la tête qu'ils qualifient :

(*quelques*/*B-NP* ***petites***/***I-NP*** *questions*/*I-NP* ***préliminaires***/***I-NP***)

Remarquons que les annotations peuvent varier selon les systèmes. Ainsi, l'annotation en chunks faite dans la campagne d'évaluation Easy, par exemple, distingue deux positions d'adjectif :

adjectif antéposé au nom : dans ce cas, il est inclus dans le chunk nominal

adjectifs postposé au nom : dans ce cas, il est la tête d'un chunk adjectival.

L'adjectif *préliminaires* dans notre exemple, ne ferait donc pas partie du chunk nominal selon ces conventions. Je ne partage pas ce choix. Selon moi, il s'agit d'un adjectif épithète quelle que soit sa position : qu'il précède le nom comme *petites* ou le suive comme *préliminaires.*

En revanche, ne sont pas compris dans les chunks nominaux les adjectifs qui ont d'autres fonctions syntaxiques que celle d'épithète détachée. Il s'agit des adjectifs apposés, ou encore d'adjectifs jouant le rôle d'attributs du sujet ou du complément. Dans ce cas, ils font partie d'un chunk adjectival :

(*c'*/*B-NP*) (*est*/*B-VN*) (***vrai***/***B-AP***)
(*c'*/*B-NP*) (*est*/*B-VN*) (*rendre*/*B-VN*) (***vivant***/***B-AP***)

- adverbes

Les adverbes fonctionnent dans un énoncé comme modifieurs soit d'un énoncé entier, et dans ce cas ils forment un chunk adverbial *AdP* :

(*y*/*B-VN* *aura*/*I-VN*) (***encore***/***B-AdP***) (*des*/*B-NP* *médecins*/*I-NP*)(***quand_même***/***B-AdP***)
(*elle*/*B-NP*) (*est*/*B-VN*) (*normale*/*B-AP*) (***maintenant***/***B-AdP***)

soit du groupe auquel ils se rattachent. C'est par exemple le cas lorsque l'adverbe modifie un adjectif et se trouve à l'intérieur d'un chunk nominal *NP* ou adjectival *AP* :

il/$_{B\text{-}NP}$ fait/$_{B\text{-}VN}$ ***trop/$_{B\text{-}AP}$*** *chaud/$_{I\text{-}AP}$*

(ces/$_{B\text{-}NP}$ enfants/$_{I\text{-}NP}$ ***très/$_{I\text{-}NP}$*** *inadaptés/$_{I\text{-}NP}$)*

Les chunks adverbiaux peuvent contenir des multi-mots comme dans l'exemple ci-dessus (*quand même*).

- conjonctions

Les conjonctions de coordination et de subordination sont annotées par le chunk approprié *CONJ*

*(je/$_{B\text{-}NP}$) (trouve/$_{B\text{-}VN}$) (****que/$_{B\text{-}CONJ}$****)*

Deux types de découpage

SEM propose deux types de découpage, l'un comprenant *tous les chunks* cités ci-dessus :

(je/B-NP) (suis/B-VN) (institutrice/B-NP) (d'/B-PP enfants/I-PP inadaptés/I-PP)

l'autre en *chunks NP* uniquement :

(je/B) (suis/O) (institutrice/B) (d'/O) (enfants/B inadaptés/I)

Dans le premier cas, à chaque chunk est associée une étiquette qui marque le début du chunk (*B*) ou sa continuité (*I*). Dans le deuxième, les mots qui ne font pas partie d'un chunk nominal sont annotés par une étiquette (*O*) désignant « Out ».

Certains chunks peuvent en contenir d'autres, comme dans le cas du chunk prépositionnel constitué des prépositions et de leurs compléments.

Dans l'exemple ci-dessus, la préposition *d'*, qui introduit un chunk PP dans le chunking complet, est annoté O dans le chunking NP. De la même manière, le groupe nominal *enfants inadaptés* forme un chunk nominal si l'on suit le découpage en NP uniquement mais il fait partie d'un chunk prépositionnel *d'enfants inadaptés* dans le cas du chunking complet.

La différence de découpage peut être aussi observée en ce qui concerne les cas de coordination. Si dans le chunking complet un chunk conjonction *CONJ* se place entre les deux éléments qu'il coordonne :

*(l'/$_{B\text{-}NP}$ école/$_{I\text{-}NP}$ publique/$_{I\text{-}NP}$) (****et/$_{B\text{-}CONJ}$****) (privée/$_{B\text{-}AP}$)*

ce n'est pas le cas dans le chunking nominal :

(l'/$_{B\text{-}NP}$ école/$_{I\text{-}NP}$ publique/$_{I\text{-}NP}$ ***et/$_{I\text{-}NP}$*** *privée/$_{I\text{-}NP}$)*

Il est possible de récupérer l'ensemble en un seul chunk, si toutefois il ne s'agit pas de deux chunks nominaux puisqu'il ne peut y avoir qu'une seule tête nominale dans un NP :

(*l'*/*B-NP* *école*/*I-NP* *publique*/*I-NP*) (***et***/***B-CONJ***) (*l'*/*B-NP* *école*/*I-NP* *privée*/*B-AP*)

Deux types de tests de chunking ont été effectués. Tout d'abord, nous avons voulu tester et évaluer les résultats du chunking du logiciel SEM appris à la base sur le corpus écrit sur nos données orales. Par la suite, nous avons essayé d'apprendre automatiquement la segmentation en chunks à partir du corpus oral annoté avec les chunks plus adaptés à l'oral, c'est-à-dire qui tiennent compte des disfluences.

Dans la partie qui suit, je présenterai le choix effectué pour les étiquettes, à savoir les adaptations effectuées pour assurer la transposition de l'écrit à l'oral. Ces adaptations et modifications portaient surtout sur les disfluences absentes dans le corpus écrit. Il fallait donc tout d'abord adapter les étiquettes POS et les chunks correspondant au corpus FTB aux cas de disfluence.

3.3.3.3.2. *Adaptation des étiquettes POS*

Le jeu d'étiquettes morphosyntaxiques utilisé par les modèles appliqués par SEM est celui de (Crabbé et Candito 2008). Tableau 3 montre la correspondance entre les cas de disfluences et les étiquettes POS de SEM.

Disfluences	**POS de SEM**	**Exemple**
euh d'hésitation	***I***	***(euh)I*** *l- dans ma classe*
interjections		*on peut commencer* ***(bon)I*** *alors*
marqueurs discursifs		*des idées laïques* ***(quoi) I***
faux départs et amorces	***UNKNOWN*** *(impossibles à interpréter)*	*euh* ***(l-)UNKNOWN*** *dans ma classe*
	étiquette selon le contexte	vous êtes **(in-)NC** institutrice
répétition	deux mêmes étiquettes	***(et)CC (et)CC*** *elle me disait*
		(la)DET (la)DET *fille*

Tableau 3 : Adaptation des étiquettes POS de SEM à l'oral

Les trois phénomènes (hésitations, interjections et marqueurs discursifs) ont reçu la même étiquette *I* (interjection). SEM a une étiquette *(UNKNOWN)* correspondant aux mots étrangers, aux néologismes dans FTB. J'ai utilisé cette étiquette pour les unités tronquées qu'on ne peut pas interpréter dans le contexte. Dans les cas inverses, si l'interprétation est possible, on annote selon le contexte. Ainsi, dans les exemples, l'amorce *in-* correspond exactement au début du mot suivant *institutrice*. On va le considérer en tant que nom commun. Dans le cas de répétition, les deux unités reçoivent la même étiquette.

3.3.3.3.3. Adaptations sur les choix de chunks

Découper en chunks la transcription de l'oral pose des problèmes spécifiques. Certains choix sont spécifiques aux disfluences (voir Tableau 4).

Répétitions

En cas de répétition, deux possibilités se présentent :

- Si l'élément répété est la tête du groupe syntaxique, il est nécessaire de distinguer les deux chunks car un chunk ne peut pas avoir deux têtes.

  (et/$_{CC}$)$_{CONJ}$ (et/$_{CC}$)$_{CONJ}$ *(elle/$_{CLS}$)$_{NP}$ (me/$_{CLO}$)$_{NP}$ (disait/$_{V}$)$_{VN}$*

- Dans le cas inverse, les deux éléments appartiennent au même chunk

  (la/$_{DET}$ la/$_{DET}$ *belle/$_{ADJ}$ jeune/$_{ADJ}$ fille/$_{NC}$)$_{NP}$*

Marqueurs discursifs

Il a été mentionné ci-dessus que les marqueurs discursifs sont considérés comme interjection par l'étiquetage POS de SEM. Les interjections ne pouvant pas être tête de chunk, ils font partie des chunks adverbiaux.

(on/$_{CLS}$)$_{NP}$ (peut/$_{V}$)$_{VN}$ (commencer/$_{VINF}$)$_{VN}$ ***(bon/$_{I}$)$_{AdP}$*** *(alors/$_{I}$)$_{AdP}$*

Disfluences	**Chunks de SEM**	**Exemple**
euh d'hésitation	***AdP***	***(euh)AdP*** *l- dans ma classe*
interjections		*on peut commencer* ***(bon)AdP*** *alors*
marqueurs discursifs		*des idées laïques* ***(quoi) AdP***
	sauf les cas où ils se trouvent à l'intérieur d'un groupe	*(l'école* ***euh*** *publique)****NP***
faux départs et amorces	***AdP*** *(impossibles à interpréter)*	*euh* ***(l-)AdP*** *dans ma classe*
	chunk selon le contexte	vous êtes **(in-)NP** institutrice
répétition	deux chunks (si tête du groupe)	***(et)CONJ (et)CONJ*** *elle me disait*
	font partie du même chunk dans le cas inverse	***(la la*** *fille)****NP***

Tableau 4 : Adaptation des chunks de SEM à l'oral

Faux départs et amorces

Les faux départs et les amorces impossibles à interpréter sont étiquetés en tant qu'interjections et font donc partie des chunks adverbiaux.

(c'/$_{CLS}$)$_{NP}$ (est/$_{V}$)$_{VN}$ (difficile/$_{ADJ}$)$_{AP}$ (euh/$_{I}$)$_{AdP}$ ***(les/$_{I}$)$_{AdP}$*** *(dans/$_{P}$ ma/$_{DET}$ classe/$_{NC}$)$_{PP}$*

Dans les autres cas où l'interprétation est possible, on annote selon le contexte :

(vous/$_{PRO}$)$_{NP}$ (êtes$_{V}$)$_{VN}$ ***(in-/$_{NC}$)$_{NP}$*** *(institutrice/$_{NC}$)$_{NP}$*

(chez/$_{P}$ vous/$_{PRO}$)$_{PP}$ ***(chez/$_{P}$ v-/$_{PRO}$)$_{PP}$***

Ainsi, dans les exemples, l'amorce *in-* correspond exactement au début du mot suivant *institutrice*. On va le considérer en tant que nom commun et il formera un chunk nominal. Dans l'exemple suivant, la répétition de la même préposition *chez* et l'équivalence entre l'amorce *v-* et le début du pronom *vous*, laisse supposer qu'il s'agit de la répétition du même groupe prépositionnel. Comme on le voit, l'interprétation concernera les deux niveaux d'étiquetage POS et chunks.

3.3.3.3.4. Résultats

Le résultat se présente sous forme de trois colonnes : token, étiquette POS, étiquette *B/I* indiquant le débit et l'intérieur du chunk repéré.

ça	*PRO*	*B-NP*
marche	*V*	*B-VN*
ou	*CC*	*B-CONJ*
quoi	*I*	*B-AdP*
oui	*I*	*B-IntP*
en_effet	*ADV*	*B-AdP*
on	*CLS*	*B-NP*
peut	*V*	*B-VN*
commencer	*VINF*	*B-VN*
bon	*I*	*B-AdP*

L'extrait du corpus oral ESLO annoté en chunks en gardant les étiquettes du FTB est présenté dans l'Annexe 8.

I	II	III	IV
Tokens	**POS proposés par SEM**	**POS corrigés à la main**	**Chunks « type FTB » corrects**
euh	*DET*	*I*	*AdP-B*
l-	*DET*	*UNKNOWN*	*AdP-B*
dans	*P*	*P*	*PP-B*
ma	*DET*	*DET*	*PP-I*
classe	*NC*	*NC*	*PP-I*

Tableau 5 : Résultats

Le Tableau 5 montre le traitement appliqué à un énoncé *euh l- dans ma classe* (colonne I). Tout d'abord, il est étiqueté en POS par le logiciel SEM (colonne II). On aperçoit deux erreurs de cet étiquetage liées à la présence de disfluences (*euh* d'hésitation et une amorce *l-*) qui ne sont pas reconnues par le système appris sur le corpus écrit et qui se trouvent étiquetées comme déterminant. Le corpus étiqueté en POS par SEM a été ensuite corrigé manuellement (colonne III). Ces deux erreurs se trouvent corrigées. Le *euh* d'hésitation reçoit une étiquette *I* (interjection) et une amorce impossible à interpréter est étiqueté *UNKNOWN*. Le corpus est enfin chunké par SEM et corrigé ensuite (colonne IV). Les deux disfluences forment deux chunks adverbiaux *AdP* selon les conventions adoptées.

L'évaluation a été effectuée deux fois.

- Dans le cas où le corpus a été étiqueté avec les POS de FTB et a été ensuite chunké par SEM, la F-mesure varie entre 77,24% et 76%.
- Dans le cas où le chunking est appliqué sur le corpus étiqueté et corrigé, la F-mesure varie entre 87,74% et 88,43%.

Les résultats de l'évaluation sont meilleurs de dix pourcent sur les étiquettes POS corrigées. Pourtant, ces chiffres sont loins de ceux de l'application de SEM sur des corpus de l'écrit (99 % sur un corpus journalistique), ce qui s'explique par le fait que SEM n'a pas été entraîné sur un corpus de même type que celui sur lequel il a été appliqué.

3.3.3.4. Apprentissage automatique à partir du corpus oral annoté

Utiliser un chunker entraîné sur un corpus écrit n'est pas idéal, ce que confirment les faibles résultats. D'où les essais pour apprendre directement le nouveau mode de chunking à partir d'un extrait de corpus oral annoté manuellement.

Disfluences	**Chunks de SEM**	**Exemple**
euh d'hésitation	***IntP***	***(euh) IntP*** *l- dans ma classe*
interjections		*on peut commencer* ***(ben) IntP*** *alors*
marqueurs discursifs		*des idées laïques* ***(quoi) IntP***
	sauf les cas où ils se trouvent à l'intérieur d'un groupe	*(l'école* ***euh*** *publique)****NP***
faux départs et amorces	***UNKNOWN*** (impossibles à interpréter)	*euh* ***(l-) UNKNOWN*** *dans ma classe*
	chunks selon le contexte	vous êtes **(in-)NP** institutrice
répétition	Deux chunks si la tête du groupe	***(et)CONJ (et)CONJ*** *elle me disait*
	sinon font partie du même chunk	*(****la la*** *fille)****NP***

Tableau 6 : Conventions propres à l'oral

3.3.3.4.1. Constitution du corpus de référence avec les nouveaux chunks adaptés à l'oral

Le corpus de référence constitué manuellement tient compte cette fois-ci des caractéristiques de l'oral.

Comme on peut voir dans le Tableau 6, la liste des chunks a été élargie au moyen de deux nouveaux chunks :

UNKNOWN

L'étiquette *UNKNOWN* existe dans le jeu d'étiquettes de FTB. Elle est attribuée aux mots étrangers. J'ai utilisé cette étiquette pour désigner :

- les erreurs de transcriptions ;
- les faux départs et les amorces dont l'interprétation est impossible.

> *(c'/$_{CLS}$)$_{NP}$ (est/$_{V}$)$_{VN}$ (difficile/$_{ADJ}$)$_{AP}$ (euh/$_{I}$)$_{IntP}$* ***(les/$_{UNKNOWN}$)$_{UNKNOWN}$*** *(dans/$_{P}$ ma/$_{DET}$ classe/$_{NC}$)$_{PP}$*

Dans cet exemple, la forme *les* est difficile à comprendre. S'agit-il d'un pronom, d'un déterminant ou d'une amorce ? L'étiquette *UNKNOWN* a été introduite aussi pour le niveau POS où se retrouve ce problème d'interprétation.

Chunk d'interjection (*IntP*)

Nous avons pu observer dans la section précédente le problème que posent des marqueurs discursifs classés dans les chunks adverbiaux. L'ajout d'un nouveau chunk *IntP* destiné à tous les marqueurs discursifs et aux interjections résout ce problème au moins partiellement. Je n'ai pas voulu ajouter un chunk supplémentaire propre aux marqueurs discursifs. Cette démarche qui réunit les interjections et les marqueurs discursifs sous la même étiquette suit le principe adopté pour l'étiquetage morphosyntaxique où l'étiquette du premier niveau *MD* regroupe les trois sous-classes (marqueur discursif *MD*, marqueur discursif interjection *MDINT* et marqueur discursif *euh* d'hésitation *MDEUH*).

> *(c'/$_{CLS}$)$_{NP}$ (est/$_{V}$)$_{VN}$ (difficile/$_{ADJ}$)$_{AP}$* ***(euh/$_{I}$)$_{IntP}$*** *(les/$_{UNKNOWN}$)$_{UNKNOWN}$ (dans/$_{P}$ ma/$_{DET}$ classe/$_{NC}$)$_{PP}$*

En revanche, les interjections et les marqueurs discursifs peuvent se retrouver à l'intérieur d'un groupe syntaxique comme dans cet exemple :

> *(l'/$_{DET}$ école/$_{NC}$* ***euh/$_{I}$*** *publique/$_{ADJ}$)$_{NP}$*

ce qui n'est pas le cas de l'exemple suivant où le marqueur discursif *quoi* est exclu du chunk nominal *des idées laïques* qui le précède :

> *(des/$_{DET}$ idées/$_{NC}$ laïques/$_{ADJ}$)$_{NP}$* ***(quoi/$_{I}$)$_{IntP}$***

Le corpus ainsi constitué contient 8093 tokens annotés en 2489 chunks (voir l'extrait de ce corpus dans l'Annexe 9).

3.3.3.4.2. Apprentissage et résultats

Expériences

Isabelle Tellier a proposé d'effectuer plusieurs expériences d'apprentissage sur les donnnées annotées. La première expérience consiste à apprendre un chunker à partir de données cibles annotées en POS corrigées (colonne III du Tableau 5) et de chunks adaptés à l'oral (Tableau 6). Nous obtenons une F-mesure variant de 96,06% à 96,65% ce qui montre que les résultats obtenus sont meilleurs que lors des expériences précédentes où l'oral n'avait pas été pris en compte.

Au cours de la deuxième expérience, le chunker est appris sur des étiquettes POS corrigées (colonnes III et V du Tableau 5) mais utilisé sur des données avec des étiquettes POS non corrigées (colonne II du Tableau 5). Nous obtenons ainsi une micro-moyenne des F-mesure de 73,81, et une macro-moyenne de 59,62.

Enfin, la dernière expérience vise à réaliser l'apprentissage par le chunker de l'oral uniquement par des étiquettes POS fournies par SEM, sans correction, ni en apprentissage ni en test. L'objectif de cette dernière expérience est d'évaluer le résultat d'un apprentissage par un bon chunker en dépit d'étiquettes POS médiocres. Nous obtenons alors une micro-moyenne de 88,84, et une macro-moyenne de 81,76. Cette expérience semble être la plus prometteuse (Annexe 9).

Analyse

Quelques remarques concernant les résultats de la dernière expérience d'annotation de nouveaux chunks. Les chunks *(IntP)* sont bien reconnus (plus de 93 de F-mesure), alors que SEM substitue à l'étiquette POS correcte *I* des étiquettes assez variées (typiquement *ADV, ADJ, NC* et *V*). Mais les interjections sont à la fois fréquentes et assez peu variées dans ce corpus oral (*euh, hm, ben etc.*) et celles présentes dans l'ensemble d'apprentissage suffisent apparemment au chunker entraîné à les identifier dès lors qu'il a accès aux mots ou tokens et pas uniquement aux POS. Sur le chunk *(UNKNOWN)*, le nouveau chunker obtient une bonne précision (92,86%) mais un mauvais rappel (18,57%). Cela tient sans doute au fait que les chunks inconnus peuvent parfois correspondre à des mots connus mais identifiés dans un contexte inapproprié. En outre, les amorces présentent une très grande variabilité contrairement aux interjections ; toutes ne peuvent pas être présentes dans l'ensemble d'apprentissage. Ces résultats peuvent être améliorés si l'on tient compte des conventions de transcription où les amorces ont été signalées par un tiret à droite suivi d'une espace. Pour une analyse détaillée des résultats concernant les différents chunks et les différentes expériences, voir (Tellier *et al.* 2013, 2014).

3.3.3.5. Conclusion

Cette série d'expériences pour chunker le corpus oral a démontré l'importance de la prise en compte des données de l'annotation. Nous avons choisi de tenir compte des propriétés différentielles de l'oral. La tâche de chunking devient plus complexe car il faut annoter huit types de chunks au lieu de six. Nous avons conservé deux contraintes : un corpus de référence de petite taille et une absence de correction manuelle des POS.

En présence d'étiquettes POS correctes et cohérentes avec les chunks, l'apprentissage automatique joue parfaitement son rôle, et permet d'obtenir l'apprentissage d'un chunker d'aussi bonne qualité que celui appliqué à l'écrit avec bien plus de données. En revanche, un tel chunker dépend fortement des étiquettes POS sur lesquelles il se fonde : l'absence de correction manuelle fait chuter ses performances. La dernière expérience montre qu'on peut

obtenir l'apprentissage à partir d'un chunker spécifique de l'oral d'assez bonne qualité (y compris pour la reconnaissance des interjections par exemple), en s'appuyant uniquement sur un petit nombre de données annotées avec des étiquettes POS de qualité moyenne.

Ce travail effectué sur un corpus non standard constitué de transcriptions met en évidence l'apport des compétences linguistiques dans le développement des outils de l'annotation automatique. Une bonne maîtrise des particularités de l'oral et leur prise en compte au niveau du jeu d'étiquettes s'intègrent au processus technique de l'annotation automatique, en convergence avec la connaissance des théories, des modèles et des classifications linguistiques existant dans le respect des contraintes liées à la structure de la langue.

3.3.4. Bilan

3.3.4.1. Compte rendu

	Etiquetage en POS	Chunking
Objectif visé	Développer un outil d'annotation automatique du corpus oral en POS et en chunks en tenant compte de sa nature	
Méthodologie	Méthode statistique fondé sur l'apprentissage automatique avec les CRF à partir d'un corpus de référence	
Données traitées	Transcriptions des entretiens en face-à-face d'ESLO1	
Difficultés rencontrées	Ambiguïté Expressions polylexicales ou multi-mots Mots mal orthographiés Disfluences et marqueurs discursifs Absence de ponctuation	
Résultats	Corpus de référence : un échantillon du corpus ESLO1 (les entretiens) constitué de 18424 tokens (1723 énoncés) annotés en POS Nouveau jeu d'étiquettes propres à l'oral et décomposables selon trois niveaux Etiqueteur morpho-syntaxique entraîné sur les données orales	Corpus de référence : un échantillon du corpus ESLO1 (les entretiens) constitué de 852 tokens annotés en 2489 chunks Nouveau jeu d'étiquettes propres à l'oral Chunker entraîné sur les données orales
Contraintes et limites	Non prise en compte des informations prosodiques Etiqueteur s'appliquant à un corpus pré-segmenté Etiqueteur ne lemmatisant pas	Non prise en compte des informations prosodiques Corpus d'apprentissage de petite taille

Originalité et apport du travail	Développement d'un outil conçu pour l'annotation des données orales Méthodologie choisie novatrice à l'époque Nouveau jeu d'étiquettes tenant compte de l'oral Proposition de l'étiquetage par niveau	Chunker de bonne qualité entraîné sur un petit nombre de données orales annotées avec des étiquettes POS de qualité moyenne Chunker qui évite l'étape de la correction manuelle de l'étiquetage morpho-syntaxique.

3.3.4.2. Perpectives et travaux en cours

Les recherches sur l'annotation syntaxique de l'oral en collaboration avec Isabelle Tellier se poursuivent avec l'utilisation de techniques d'apprentissage automatique qui respectent au mieux la nature spécifique des données traitées. Notre problématique se formulerait de la façon suivante : « Comment peut-on annoter des corpus de types différents sans être obligé de constituer à chaque fois le corpus de référence adapté au type de corpus traité ? ». Ce problème s'est posé dans le travail sur ESLO. Pour développer un étiqueteur et un chunker appropriés, nous avons dû constituer manuellement le corpus de référence. Vu la difficulté de la tâche, le corpus constitué était de petite taille ce qui a influé sur la qualité de l'apprentissage automatique. Une autre méthode est l'adaptation d'un annotateur de l'écrit à l'oral. Nous envisageons de tester la démarche inverse que nous considérons comme plus économique et qui correspond mieux aux spécificités du corpus traité. L'idée est de partir d'un corpus écrit annoté de grande taille (FTB) et de le rapprocher au maximum de la transcription de l'oral. De cette manière, le corpus de référence sera de plus grande taille et permettra un apprentissage plus efficace de l'annotation.

Pour effectuer cette transformation, une réflexion est menée sur les types de modifications à opérer dans le corpus écrit. Tout d'abord, il convient d'effacer la ponctuation et les signes typographiques qui marquent la segmentation de l'écrit. En deuxième lieu, la réflexion doit être menée quant aux expressions multi-mots surreprésentées dans le discours oral comme les présentatifs *c'est, il y a, on a etc.* Une différence cruciale entre l'oral et l'écrit tient à la présence des disfluences. Si l'on veut rapprocher l'écrit et l'oral, il faudrait intégrer les disfluences dans le discours : répéter les déterminants, insérer les *euh* d'hésitation, les interjections et les marqueurs discursifs, prévoir les amorces etc. Les modifications peuvent porter aussi sur des différences stylistiques comme les parenthèses à l'écrit qui pourraient être remplacées par des reformulateurs paraphrastiques du type *disons, c'est-à-dire, j'allais dire, par exemple etc.* De même, l'extrait du corpus ESLO étudié reprend des entretiens en face-à-face qui portent sur l'identité du locuteur, la vie à Orléans etc. Les variations lexicales sont donc à prévoir. Ainsi, les termes comme *marché commun, évolution économique* etc. très usités dans FTB qui est composé d'articles journalistiques peuvent être remplacés par les syntagmes plus fréquents en situation d'enregistrement. La différence dans les tournures syntaxiques à l'oral et à l'écrit peut concerner par exemple les propositions relatives beaucoup moins présentes à l'oral. Les transformations doivent enfin concerner les symboles mathématiques comme *%, =, + etc.* qui sont écrits littéralement dans les transcriptions de l'oral *pourcent, égal à, plus etc.* Cette remarque concerne aussi les chiffres.

3.4. Repérage et annotation de l'information personnelle sur le locuteur

Mon travail sur le corpus ESLO a été poursuivi dans le cadre du projet ANR Variling. Il concernait l'anonymisation automatique du corpus afin d'en permettre une mise à disposition large. La réflexion sur la mise en œuvre de cette procédure m'a poussée vers un questionnement sur les différentes stratégies d'identification. Comment est-ce qu'on peut identifier le locuteur ou toute autre personne mentionnée dans le discours de celui-ci? A partir de quels indices ? Ces indices sont-ils repérables automatiquement ?

Le travail présenté est le premier dans la série des travaux sur le repérage et l'analyse de l'information sémantique. Il touche aussi une autre notion abordée dans la plupart de mes travaux, celle de l'expression de la subjectivité dans le discours.

3.4.1. Résumé du travail

L'anonymisation est une étape primordiale dans la diffusion du corpus oral selon les techniques actuelles qui impliquent une démarche fondée sur de « bonnes pratiques » juridiques et éthiques (Baude *et al.* 2006). Il ne s'agit pas de rendre totalement impossible l'identification d'un locuteur car il faudrait alors brouiller la voix sur l'ensemble de l'enregistrement, ce qui rendrait toute analyse linguistique impossible. L'idée est de repérer dans les enregistrements des éléments sensibles pouvant donner une information personnelle sur le locuteur ou toute autre personne mentionnée dans le discours.

Avant de procéder au développement du module de l'anonymisation automatique, une réflexion sur la nature des données permettant l'identification de la personne a été lancée. Elle a abouti à une définition d'une nouvelle notion : *entité dénommante*. Le travail complémentaire entrepris depuis sur la définition de cette notion m'amène maintenant à préférer la terminologie de *faisceau d'indices*.

Le travail a été effectué en collaboration avec l'équipe du Laboratoire Informatique de Tours (LI), plus particulièrement avec Denis Maurel et Nathalie Friburger. Nous avons opté pour la méthodologie de l'annotation en surface en utilisant une méthode symbolique consistant à développer des grammaires locales sous forme de graphes. Cette fois-ci, nous avons utilisé un outil développé pour l'écrit qui a été adapté à notre corpus. Nous avons procédé en deux étapes. Tout d'abord, nous lançons des cascades de transducteurs qui repèrent et annotent les entités nommées (EN). Ensuite, une autre série de cascades appliquée à ce corpus annoté, identifie les indices identifiants (DE).

Plusieurs résultats ont été obtenus grâce à cette expérience et ses apports sont multiples. Les 112 entretiens face-à-face d'ESLO1 ont été annotés en entités nommées et indices d'identification. La nature des éléments permettant l'identification de la personne a été étudiée. Cette étude a permis de développer la typologie de ces éléments que nous avons utilisée pour leur annotation. Une grammaire locale (une série de graphes) permettant la reconnaissance et le balisage automatiques des informations personnelles sur le locuteur ou sur toute autre personne mentionnée dans le discours a été créée. Le module développé ne s'arrête pas à la reconnaissance des entités nommées ce qui est le cas de nombreux travaux en TAL sur l'anonymisation (section 3.4.2) mais va plus loin en annotant d'autres éléments qui permettent au même titre que les entités nommées d'identifier le locuteur. Le travail effectué porte sur le corpus oral ce qui le distingue aussi des autres recherches dans le domaine du TAL qui traitent de préférence des corpus écrits. Enfin, l'annotation du corpus en indices d'identification a permis de faire l'analyse des différentes stratégies d'identification qu'on utilise dans le discours. Aujourd'hui, où les recherches sur l'analyse des sentiments et des opinions se multiplient, ce travail, très particulier, ajoute un aspect complémentaire à l'étude de la subjectivité dans le langage. L'anonymisation des données et la réflexion sur leur nature s'inscrit également dans les préoccupations d'aujourd'hui concernant l'éthique scientifique et le TAL y occupe une place importante.

Les travaux sur l'anonymisation du corpus ESLO et sur l'analyse des éléments permettant l'identification de la personne dans les transcriptions de l'oral ont été décrits dans (Eshkol *et al.* 2010, Eshkol-Taravella *et al.* 2012, 2015, Maurel *et al.* 2009).

3.4.2. Anonymisation du corpus

L'anonymisation est une pratique qui répond à un impératif juridique précis. Sans recueil du consentement de la personne enregistrée, il est obligatoire d'empêcher son identification. Dans le cas du corpus ESLO1, le recueil du consentement pose deux problèmes. Premièrement, il n'existe aucun document écrit par les locuteurs qui permettrait d'exprimer formellement leur accord ; deuxièmement, il serait illusoire de penser que les locuteurs de la fin des années soixante aient pu imaginer les types d'exploitation actuels, en particulier la diffusion instantanée par Internet. Le choix de l'équipe a néanmoins été d'anonymiser l'ensemble des données d'ESLO1 et d'ESLO2.

L'impossibilité d'identifier est une notion complexe, il peut s'agir d'une forme nominative, d'une profession, d'un statut, d'une caractéristique physique etc. et/ou du recoupement de plusieurs de ces informations. Il est donc nécessaire de définir avec précision quels sont les traitements à effectuer pour répondre à l'objectif de réduire les possibilités d'identification.

Traditionnellement la tâche d'anonymisation dans le TAL est décomposée en deux étapes : le repérage des entités nommées (noms de personnes, lieux, organisations, âges etc.) (Ehrmann 2008, Nadeau, Sekine, 2009) et leur substitution par un hyperonyme ou un élément à référents multiples (le nom de famille *Dupont*, par exemple). L'anonymisation dans le domaine du TAL concerne aussi souvent le domaine médical (Meyster *et al.* 2010, Tweit *et al.* 2004, Raaj 2012, Uzuner *et al.* 2007, Grouin et Zweigenbaum 2011) et porte sur des documents écrits (rapports, dossiers médicaux etc.) où les informations à anonymiser sont assez homogènes et présentées d'une manière linéaire propre aux textes écrits. C'est le cas de l'outil Medina[32] (Medical Information Anonymization) disponible gratuitement qui repère automatiquement à l'aide de patrons et de lexiques les noms de personnes, les lieux, les noms d'hôpitaux et les informations numériques comme les adresses, âges, numéros de téléphones etc. dans les documents cliniques en français.

Les corpus oraux sont différents des corpus écrits car ils n'utilisent pas un seul support mais associent le plus souvent la parole enregistrée à une représentation écrite et/ou codée (transcriptions, traductions, annotations). Le discours oral est caractérisé par la présence de multiples disfluences (hésitations, amorces, répétitions, reformulations, etc.), pauses ou chevauchements entre les locuteurs qui rompent le flux de la parole. L'information personnelle sur le locuteur ou sur une autre personne mentionnée peut apparaître n'importe où dans le corpus et de la façon « inattendue ». Ainsi, anonymiser le discours d'une manière automatique est une tâche difficile qui pose des problèmes supplémentaires au TAL comme cela a récemment été constaté (Amblard et Fort 2014). Les auteurs présentent entre autres le processus d'anonymisation automatique du discours transcrit de schizophrènes. Ils notent l'insuffisance du simple repérage à l'aide de scripts Python des mots commençant par une majuscule dans les extrais du corpus lorsque des sujets relatent un événement « s'inscrivant dans une temporalité et une géographie particulière » et la présence d'autres indices selon lesquelles on peut identifier le locuteur ou ses proches.

L'anonymisation ne se réduit donc pas à l'effacement des noms propres. Cette tâche est bien plus difficile, mais aussi plus stimulante pour les recherches en linguistique et en TAL.

[32] https://medina.limsi.fr/

3.4.3. Notion de faisceau d'indices

3.4.3.1. Traits caractéristiques de l'individu

Il est difficile de rendre compte du mécanisme cognitif en jeu dans le processus de reconnaissance d'un individu. La reconnaissance de la personne semble passer par la connaissance de certaines de ses propriétés caractéristiques. On peut supposer qu'un indice (*nom rare, handicap, caractéristique particulière*) ou une série de ces indices (*nom, métier, lieu de travail, loisir etc.*) sont associées à un individu en particulier à l'aide d'un certain lien dénominatif qui se trouve réactivé lors de leur apparition en discours. Il importe de prendre en considération les facteurs contextuels qui entourent l'énonciation de ces indices. C'est le contexte qui permettra de réduire le champ d'application de ces éléments à un seul porteur, de le distinguer des autres référents possibles comme dans le cas d'utilisation des noms propres au lieu du prénom, ou de l'anthroponyme seul, ou d'une description de l'individu. L'identification peut se faire grâce aux connaissances que l'utilisateur du corpus maîtrise concernant le locuteur ou la personne mentionnée dans son discours.

Plusieurs études linguistiques traitent des traits caractéristiques et définitoires d'un objet. En sémiotique narrative, (Hamon 1977) utilise le terme de « qualification différentielle » pour la série de traits indicatifs de l'importance des personnages dans un roman (descendant d'une famille représentée dans le même roman, porteur d'un nom propre, description physique et psycho-sociologique etc.) qui distinguent ce personnage des autres. Ces derniers ne doivent pas se voir attribuer ces propriétés ou seulement partiellement. Ce fait permet à ce personnage d'être considéré comme le héros.

En linguistique cognitive, (Jonasson 1994:141) déclare que « la description des particuliers […] comporte l'insertion de l'item dans une taxinomie, suivie de la spécification des propriétés typiques permettant son identification et éventuellement accompagnée de l'indication des associations culturelles ».

(Charaudeau et Maingueneau 2002), à leur tour, distinguent les deux types d'identité du sujet : une identité psychosociale consistant en traits qui définissent le sujet selon son âge, son sexe, son statut etc. et une identité discursive du sujet énonciateur « qui peut être décrite à l'aide de catégories locutives, de modes de prise de parole, de rôles énonciatifs et de modes d'interventions » (p.300).

3.4.3.2. Identification à travers les noms propres

Traditionnellement, ce sont les noms propres qui ont en charge de désigner une identité dans le discours. (Kripke 1972) affirme que les noms propres ne sont pas descriptifs, qu'ils ont une fonction de désignation et d'identification pures. Cette thèse est critiquée par d'autres études linguistiques.

(Jonasson 1994, Leroy 2004) ont montré qu'il est difficile de distinguer nettement le nom propre du nom commun, quels que soit les critères adoptés (usage de la majuscule, impossibilité de traduction, absence de l'article, incompatibilité avec des déterminants, mono-référentialité, manque de sens etc.). Jonasson (1994:138) affirme que le nom propre ne peut pas être considéré seulement dans son emploi référentiel. Elle distingue les noms propres « connus », « historiques », des noms propres « familiers » comme *Paul, Marie etc.* qui « sont souvent associés à de nombreux particuliers […] mais désignent dans un champ restreint (famille, classe, bureau, études, vie privée, village etc.) un seul ou un nombre limité de particuliers. » (Jonasson 1994) et (Kleiber 1981) mentionnent la difficulté d'interpréter ces noms sans ajout d'autres renseignements sur la relation personnelle, par exemple, entre le

locuteur et le référent visé (*Paul, c'est mon fils*). On a donc besoin d'une certaine extension du contexte pour pouvoir leur associer un référent. (Schnedecker 2011) mentionne un autre type de noms propres sans référents identifiables tels que *Machin, Trucmuche* etc. L'auteure les nomme « noms propres indéfinis » et note que, par leur comportement, ils sont plus proches des pronoms indéfinis que des noms propres.

Mes observations vont dans le même sens. Je partage l'idée que les noms propres n'assurent pas seulement la désignation efficiente d'un référent du monde. En outre, l'acte de désignation dépend fortement du contexte de l'énonciation. La prise en compte du contexte s'avère encore plus importante dans le cas du dialogue. Un nom de lieu, par exemple, n'a pas la même valeur s'il est présent dans la réponse sur les origines du locuteur ou s'il se trouve dans la réponse à la question sur les lieux où on parle bien le français. De la même façon, un toponyme faisant partie du nom de l'institution (*Collège de France*) ne renvoie plus vers un lieu. Enfin, comme l'a mentionné (Leroy 2004), les noms communs peuvent également désigner un référent. Il ne s'agit pas seulement d'une description définie : *le boucher vient tout à l'heure* (si l'on est dans un petit village qui n'a qu'un boucher), mais aussi d'une série de descripteurs qui, dans un contexte donné, peuvent conjointement, partiellement ou en combinaison avec des noms propres, permettre l'identification du locuteur. Ainsi, le repérage du référent se fait à partir de divers types de connaissances (linguistiques, métalinguistiques et encyclopédiques) qui, d'une manière ou d'une autre, relèvent toutes du contexte, puisqu'elles sont supposées être présentes chez les sujets parlants au moment de l'énonciation.

La désignation d'une personne ou d'un lieu est un processus social réapproprié subjectivement. En désignant quelqu'un par son prénom ou par son statut (*Madame*), on ajoute une information sur ses origines ou son statut civil. Ainsi, les noms propres peuvent aussi décrire un référent du monde. La référence à une personne ou à un lieu donnés est déterminée par les conceptions personnelles, et donc subjectives, du locuteur. En outre, lorsqu'un élément dans le discours permet d'identifier le locuteur, il n'est plus neutre car il est perçu par l'interlocuteur/auditeur d'une certaine manière. La notion d'*entité nommée* (dont les noms propres font partie) peut donc être rapprochée de celles de subjectivité qui lui est intrinsèquement liée, et de perception. On peut ainsi observer de quelle faççon on passe de l'entité nommée à une entité à connotation subjective. Cet emploi des entités nommées est conforme à la nature du corpus analysé, des entretiens oraux. On note à nouveau les effets de de la nature du corpus sur le type de traitement des données.

3.4.3.3. Faisceau d'indices dans le corpus

Le processus d'identification peut être lié directement avec celui de désignation. Selon les dictionnaires, désigner c'est « indiquer de manière à faire distinguer de tous les autres par un geste, une marque, un signe » (*Le Nouveau Petit Robert*, 1993). Dénommer c'est « attribuer un nom à quelqu'un ou à quelque chose » (*TLF* en ligne).

Pour Kleiber (1984 : 80), l'acte de dénomination « consiste en l'institution entre un objet et un signe X d'une association référentielle durable » codée, apprise, mémorisée préalablement. Il peut s'agir des noms propres comme des noms communs. La désignation, à son tour, est constituée par le processus qui crée une association occasionnelle entre un signe linguistique X et un élément de la réalité. Elle n'est donc ni codée, ni mémorisée.

Mon objectif est d'étudier des éléments dans le discours du locuteur permettant son identification par un éventuel utilisateur du corpus. Ces indices servent à identifier la personne en le mentionnant par son nom ou en représentant certains de ses traits ou de son quotidien, c'est-à-dire ce sont les éléments descriptifs qui permettent de distinguer la personne

des autres et, par conséquent, de la reconnaître. Pour identifier le locuteur ou toute autre personne dans le discours, il suffit de les nommer (s'il s'agit d'un nom rare) et/ou de les décrire par certaines de ses caractéristiques.

Le processus d'identification peut être direct ou indirect, d'où la distinction que je propose entre :

Identifiant direct (unicité référentielle) : il permet, à lui seul, de distinguer un individu des autres et renvoie directement vers un référent unique ; sa présence est suffisante pour la reconnaissance de l'individu. Le processus n'est pas progressif, il est ponctuel :

- nom rare de la personne, surnom : *Eshkol* (dans le cas de la ville comme Orléans)
- métier rare (*général*) ou statut (*maire*)
- caractéristique rare (*nombre élevé d'enfants, handicap*)

Identifiant non direct : sa présence seule ne permet pas l'identification, mais en combinaison avec d'autres identifiants, il peut désigner un référent unique : le locuteur est patron d'un bar au moment de l'enregistrement, et avant il travaillait dans l'aviation militaire. Il s'agit d'attributs qui ne sont pas uniques et peuvent être partagés par plusieurs individus. Le processus d'identification est progressif, il se construit au fur et à mesure de l'accrétion des indices.

Parmi ces identifiants non directs, on peut distinguer ceux qui sont les plus sensibles à l'anonymisation et qui apportent une information plus importante et plus spécifique, de ceux qui sont plus généraux. Cette distinction me permet d'opposer

- les noms de famille *Dupond* ou *Durand* aux autres ;
- les noms de métiers *enseignant* à *professeur de physique*

Le corpus d'entretiens « face-à-face » est un corpus riche d'informations personnelles car il est composé d'entretiens où une partie du questionnaire porte sur l'identité du locuteur : son travail, ses études, ses loisirs, sa famille. Ce choix m'a permis de travailler plus facilement sur le processus d'identification grâce à la richesse d'informations de ce type dans le corpus. Ainsi, lorsque le locuteur parle de la profession qu'il exerce, il emploie différentes formes :

- le nom direct du métier ou son synonyme : *professeur des écoles* ou *institutrice,*
- le contenu, la description de son travail : *j'enseigne les maths,*
- le lieu : *je travaille au collège de Saint-Jean-de-Braye.*

Il est très difficile de créer une typologie de ces éléments qui sont de nature très hétérogène.

Tout d'abord, il s'agit des entités nommées « classiques » repérables automatiquement. Il s'agit de « toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus » (Ehrman 2008) :

- noms de personne :

  patronnée par Suzanne Fouché

- noms de lieu :

  dans l'Indre

- noms d'organisation :

  euh je suis je travaille à l'hôpital d'Orléans quoi
  je fais partie de la SPA

- éléments chiffrés : âge, année

en mille neuf cent soixante-neuf
j'avais une fille de quinze ans

- noms de métiers

je suis enseignant dans dans l'école publique

Malheureusement, la reconnaissance des entités nommées ne suffit pas à repérer toutes les informations concernant la personne. Les entités nommées présentes dans le discours doivent avoir un lien avec le locuteur ou la personne qu'il mentionne dans son discours pour devenir un indice d'identification. Ce lien est souvent exprimé dans le discours même par le contexte gauche/droite de l'entité ou par la question posée dans le cadre de l'entretien ce qui est le cas du corpus étudié. Par conséquent, tous les noms propres ne doivent pas être anonymisés : la *Loire* et *Jeanne d'Arc* ne sont pas à inclure dans l'effacement, ainsi que les toponymes donnés en la réponse à la question *Où parle-t-on bien le français ?* ou encore le nom des animateurs célèbres de l'époque, dans les réponses sur les questions concernant les émissions télévisées ou radiophoniques. En outre, le discours du locuteur contient souvent d'autres éléments permettant de l'identifier par recoupement :

- maladies

ça lui a même occasionné une petite scoliose déformation légère de la colonne vertébrale

- études

je suis licencié licencié en physique

- loisirs

je suis scout de France le jeudi soir où j'anime un un atelier photos

- événements

mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville

etc.

Cette catégorie des indices est large. Elle inclut des éléments assez hétérogènes désignant les différentes informations personnelles sur la personne : événements, activités sociales, loisirs, maladies, handicap etc. qui peuvent au même titre que le travail, la famille donner les informations sur le locuteur ou la personne dont on parle.

Les thèmes abordés dans le cadre de l'entretien : origine, travail, études, famille, âge, loisirs etc. peuvent se recouper :

actuellement j'enseigne à côté de Châteauroux et j'étudie à Orléans

mon fils est venu manger qui est soldat

Chaque type d'information peut être présenté à travers un groupe nominal ou des expressions plus étendues. Ce passage se manifeste par l'ajout de propriétés supplémentaires à la classe présentée par le groupe nominal minimal, ce qui diminue l'extension de la classe et rapproche le groupe d'une référence plus individualisante. Prenons comme exemple le domaine du travail, lorsque le locuteur essaie de le décrire et de donner plus de détails sur ses fonctions :

je suis professeur d'éducation physique

j'enseigne l'éducation physique dans toutes les classes de la sixième à la troisième

je dispose d'un d'un plateau d'éducation physique qui comporte deux terrains de basket un terrain de hand-ball

je suis au collège de Saint-Jean-de-Braye

mon travail d'élève infirmière ou le travail d'auxiliaire de puériculture

je m'occupe uniquement de malades adultes puisque je fais mes études d'infirmière

Le locuteur nomme d'abord son métier (à cause des questions posées) et ensuite le spécifie.

Suite à ces exemples précis, on constate que la spécification du travail s'effectue souvent par :

- les verbes d'activité : *s'occuper de, faire de, enseigner* etc. + domaine d'activité
- une précision concernant le lieu d'exercice : entité nommée introduite par *être* avec fonction locative ou par une préposition locative

Ainsi, on peut observer une certaine régularité des structures syntaxiques (constructions attributives, verbes d'occupation, noms de relation familiale etc.), ce qui rend possible la création de patrons pour une reconnaissance automatique. Cependant, on constate la présence d'informations occasionnelles « imprévisibles », donc non homogènes à travers le corpus, comme :

nous louons une villa à Royan

mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville

j'attends un deuxième bébé

En conclusion, le faisceau d'indices inclut les entités nommées identifiantes, mais peut contenir aussi d'autres éléments qui permettent l'identification soit directement, soit, par combinaison au sein de ce faisceau.

3.4.3.4. Pertinence d'un indice

La multitude d'éléments personnels, en particulier biographiques, dans le corpus soulève la question de leur pertinence. Sont-ils tous également identifiants ? Lesquels faut-il retenir comme pertinents pour l'anonymisation ? Est-ce que *l'année d'arrivée à Orléans, l'âge des enfants, le lieu de naissance* (autre que *le Loiret*), *la nationalité* (si elle n'est pas *française*) etc. sont susceptibles de révéler l'identité du locuteur ou de ses proches ?

En observant le corpus, on constate que c'est souvent le recoupement de plusieurs indices qui permet de lever l'identité de la personne. Être un *professeur* ne permet pas d'identification, mais il n'en va pas de même s'il est précisé par ailleurs que c'est un *professeur d'université spécialisé en électronique* et, ailleurs encore, que c'est une *femme*, auquel cas on peut arriver à un singleton. À cela s'ajoutent des exemples comme *mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville* qui sont peu présents dans le corpus mais permettent une identification immédiate. Il est nécessaire ainsi de faire une distinction entre les informations personnelles sur le locuteur et les informations personnelles identifiantes. L'annotation de ces derniers dépend du contexte pris au sens le plus large et reste en partie une opération subjective.

3.4.4. Balisage

3.4.4.1. Méthodologie adoptée

L'annotation des entités nommées et des indices d'identification a été décrite dans (Eshkol-Taravella *et al.* 2012, 2015, Eshkol *et al.* 2010, Maurel *et al.* 2011, 2009). Je présente ici une synthèse de ces articles.

Pour repérer et annoter les indices d'identification, nous avons choisi l'approche en surface permettant de construire les grammaires locales selon le contexte en utilisant le système CasSys (Friburger 2002) intégré à la plate-forme Unitex (Paumier 2003). Ce choix nous a évité le développement d'un outil par l'adaptation d'un outil existant à nos données, ce qui étatit plus économique et approprié à la tâche. L'analyse du corpus a permis de créer des règles d'extraction (patrons) fondées sur des questions et sur les structures répétées dans les réponses.

Le choix du corpus s'est porté sur un sous-corpus d'ESLO1[33] : les entretiens en « face-à-face ». Ils contiennent nombre d'informations personnelles puisque le questionnaire inclut comporte des interrogations telles que : *Depuis combien de temps habitez-vous Orléans ? Quel âge avez-vous ? Qu'est-ce que vous faites comme métier ? Où travaillez-vous ? Qu'est-ce que fait votre époux(se) ?* etc.

Ce corpus présente bien des avantages. Tout d'abord, les réponses des locuteurs montrent comment les Orléanais de l'époque parlent d'eux-mêmes. A ce titre, il représente une masse de données valorisant une variation intéressante à analyser du point de vue linguistique mais aussi sociologique. Le choix des entretiens pour l'annotation automatique s'explique aussi par l'homogénéité et la richesse des données personnelles. Dans les discours spontanés, les énoncés contenant des informations personnelles permettant une identification sont plus rares et surtout moins structurés, ce qui rend plus difficile une annotation automatique.

112 entretiens en face-à-face ont été sélectionnés. L'annotation a été réalisée sur les fichiers de transcription Transcriber, soit un total de 35,75 Mo. Six fichiers ont été réservés pour les tests et neuf pour l'évaluation.

3.4.4.2. Repérage et balisage des entités nommées

La cascade CasEN[34] utilisée a été conçue à l'origine 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 :

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>

oh j'ai une <ENT type="prod.art"> encyclopédie Quillet </ENT>

[33] À l'origine de cette étude les transcriptions ESLO2 n'étaient pas disponibles.

[34] La cascade utilisée pour les entités nommées est disponible sous licence LGPL-LR à l'URL : http://tln.li.univ-tours.fr/Tln_CasEN.html

3.4.4.3. Repérage et balisage des indices d'identification

Nous avons créé une nouvelle cascade appelée *CasDen* ayant pour finalité le repérage de toute information personnelle (famille, travail, engagement...). Certaines de ces informations ont servi à l'anonymisation du corpus, d'autres pourront permettre des études sociologiques sur la vie à Orléans durant cette époque. Cette nouvelle cascade s'applique sur le texte balisé par la cascade *CasEN*. Un extrait du corpus annoté par la cascade *CasDen* est présenté dans l'Annexe 11.

Elle étiquette les entités nommées identifiantes, c'est-à-dire ayant un certain lien avec le locuteur ou la personne mentionnée. En outre, l'entité nommée repérée doit être étiquetée selon son rapport avec le locuteur. Dans la phrase *je travaille au collège de Saint-Jean-de-Braye*, l'entité *collège de Saint-Jean-de-Braye* n'est plus seulement un établissement scolaire mais également le lieu de travail du locuteur. Le processus d'anonymisation ne s'arrête pas à la reconnaissance des entités nommées, car il s'agit aussi du repérage des éléments d'identification hors noms propres.

La méthodologie est fondée sur la prise en compte du contexte. En premier lieu, on peut mentionner le contexte immédiat (gauche et/ou droit) d'un indice. Ainsi, un nom de lieu employé seul ne présente guère d'intérêt, mais avec des verbes comme *venir de, travailler à* ou avec des noms comme *collège, hôpital* etc. il devient un indice du lieu de travail, d'études ou d'origine de la personne. Les patrons décrivent le syntagme et son contexte immédiat en utilisant des marqueurs lexicaux (mots déclencheurs), des dictionnaires de noms propres et des dictionnaires spécifiques (par exemple un dictionnaire des métiers). Ces indices permettent de repérer un élément mais aussi de le catégoriser. Un autre contexte utilisé dans cette approche est celui de la question posée. On outrepasse dans ce cas les limites de l'énoncé pour étudier un contexte plus large. Le nom de lieu, par exemple, n'est pas signifiant s'il est utilisé pour répondre à la question *Où parle-t-on le mieux le français ?* par contre il devient un indice identifiant dans les réponses aux questions concernant les origines du locuteur, ou dans les énoncés décrivant le métier du locuteur, pour autant que celui-ci indique le lieu de son travail. De la même manière, les réponses aux questions sur les émissions de télévision, par exemple, n'apportent pas d'information personnelle et les noms de personnes qui apparaissent n'ont pas à être pris en compte :

monsieur Fouchet Christian Fouchet ministre de l'Education Nationale

il a surpris beaucoup de personnes Edgar Faure certainement

Une distinction a été ainsi proposée entre les questions sensibles dont les réponses peuvent contenir certaines informations personnelles et qui ont été prises en compte par les graphes :

Qu'est-ce que vous faites comme travail ?

En quoi est-ce que ça consiste / c'est quoi au juste ?

Et votre femme, est-ce qu'elle travaille aussi? Pourquoi (pas) ?

Et vos enfants, que font-ils ? / Leur métier ?

Qu'est-ce que vous faites de votre temps libre – soirées, week-end ?

Comment a-t-on choisi dans votre cas personnel entre l'école publique et l'école libre ?

etc.

et les questions neutres où la présence des entités nommées ne renvoie pas nécessairement au locuteur :

> *A votre avis, qu'est-ce qu'on devrait apprendre surtout aux enfants à l'école ?*
>
> *Qu'est-ce que vous pensez du latin à l'école ?*
>
> *Pour revenir à la ville d'Orléans, est-ce que, d'après vous, on fait assez pour les habitants d'Orléans?*
>
> *Ecoutez-vous la radio ? Le nombre d'heures par semaine/jour ? Votre chaîne préférée ? Vos émissions préférées ?*
>
> *etc.*

Enfin, nous avons pris en compte le contexte socioculturel de l'époque ce qui est le cas des destinations de vacances pour le corpus ESLO1 car en 1968 très peu de gens voyageaient à l'étranger :

> *nous sommes allés par bateau jusqu'au Cap Nord et retour euh par euh jusqu'à la frontière finlandaise jusqu'à Oslo après nous avons vu euh la Suède et le Danemark Canaries et retour par Dakar*

Pour conserver une certaine homogénéité entre les deux cascades, une typologie en six points a été définie :

1) *personne* repère les informations sur la personne interrogée et celles sur sa famille ;

2) *identité* marque des informations précises comme la date de naissance ou la date d'arrivée à Orléans, l'âge de la personne dont on parle, son origine, la date de son mariage etc. ;

3) *travail* étiquette le métier, le secteur d'activité, le lieu de travail ou le nom de l'entreprise de la personne dont on parle ;

4) *engagement* concerne la vie associative (y compris syndicale ou parentale) et militaire ;

5) *voyage* répertorie les différents déplacements, plus rares à cette époque qu'aujourd'hui ;

6) *études* indique les diplômes, les lieux d'apprentissage ou les établissements d'enseignement.

Pour une présentation, voir l'Annexe 10.

Voici ce que donne le balisage suite au 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>*

Ainsi, nous annotons tout d'abord la personne qui donne l'information : le locuteur ou les autres membres de sa famille ; nous précisons ensuite la nature de cette information : l'identité, le travail, les études, l'engagement associatif ou syndical, les vacances.

Pour évaluer la cascade des indices d'identification, nous avons utilisé 9 enregistrements (les 6 fichiers ont été réservés pour les tests)[35]. Sur les 77 éléments personnels, nous en avons reconnu 69, il y a eu 4 erreurs et 12 oublis. La précision est de 94,2 % et le rappel de 84,4 %.

Parmi les éléments non reconnues, certains étaient dues aux disfluences (*euh j'habitais dans dans le <ENT type="loc.admi"> Berry </ENT> à <ENT type="loc.admi"> Bourges </ENT>*), d'autres à des oublis dans la cascade des entités nommées (*je travaille actuellement à l'agence financière du <ENT type="loc.geo">bassin Loire-Bretagne</ENT>* où la présence d'*agence financière* aurait dû permettre le balisage d'une organisation et donc celui d'une entreprise où travaille le témoin).

3.4.4.4. Difficultés du travail effectué

La reconnaissance des informations personnelles se distribue entre plusieurs groupes syntaxiques, qui peuvent relever de plusieurs locuteurs. Il était indispensable de prévoir la présence éventuelle de disfluences. Deux sous-graphes spécifiques ont été écrits pour détecter respectivement les pauses simples, les insertions et amorces. Les répétitions ne sont pas traitées et leur correction a été manuelle. Opérant sur des fichiers de transcription issus de Transcriber, il fallait prendre en compte les balises XML. Certains graphes utilisent la question comme amorce. Ils comportent une description du découpage XML de Transcriber avec la balise {S} qu'utilise Unitex pour segmenter le document analysé. Par exemple le graphe de la Figure 2 comporte quatre sous-graphes décrivant respectivement une question (sur la date d'arrivée à Orléans), les balises XML Transcriber et la segmentation, une disfluence éventuelle et la réponse à la question.

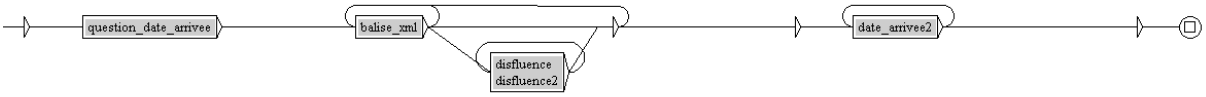

Figure 2 : Un graphe question-réponse sur la date d'arrivée à Orléans

Certaines indications ne résultent pas d'une question. Elles se trouvent disséminées ici ou là dans la conversation. Par exemple la description de la recette de l'omelette peut donner l'occasion d'introduire son origine géographique :

> *comment qu'on fait une omelette ?*
> *…*
> *enfin on assaisonne sel poivre euh nous en Lorraine on on découpe des petits des petits morceaux de lards qu'on fait frire avan*t

Le contexte socioculturel de l'époque mentionné dans la section précédente demande des connaissances extralinguistiques et ne peut donc pas être totalement modélisable. De même pour certaines informations déduites du contexte comme dans l'exemple suivant :

> *BV: y a longtemps que vous êtes à Orléans ?*

[35] Une évaluation de cette cascade est présentée dans (Maurel *et al.*, 2009)

MS530: euh oui euh vingt-deux ans

BV: ça fait euh vous êtes née à Orléans

MS530:oui

La prise en compte du contexte socioculturel de l'époque et la déduction de nouvelles informations montrent les limites du traitement automatique sur des corpus qui ne correspondent pas à la situation présente. La difficulté réside dans la description formelle de ce type de connaissances qui fait partie des recherches en intelligence artificielle, mais qui se rencontre moins souvent dans le domaine des corpus oraux.

A cela s'ajoutent des informations occasionnelles « imprévisibles », et donc non homogènes, sur le locuteur qui sont difficilement repérables automatiquement :

on a monté une association d'élèves infirmières

nous louons une villa à Royan

mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville

j'attends un deuxième bébé

Les informations de nature personnelle varient d'une manière non homogène dans le corpus. Ainsi, le locuteur peut décrire son métier de manières diverses : *je suis enseignant dans l'école publique, je suis maître auxiliaire, j'enseigne des mathématiques modernes des mathématiques classiques de la chimie et de la technologie etc.* On ne peut donc jamais atteindre une liste exhaustive de toutes les reformulations possibles.

3.4.5. Etape finale d'anonymisation

Pour anonymiser les entretiens d'ESLO1, les indices qui renvoient vers les informations personnelles concernant le locuteur et sa famille et qui peuvent éventuellement permettre sa reconnaissance, ont été repérés et étiquetés. Ils ont été validés manuellement. Ceux qui identifient directement le locuteur ou toute autre personne mentionnée dans le corpus ont été remplacés par un hyperonyme.

3.4.6. Bilan

3.4.6.1. Compte rendu

Objectif visé	Faire un test de l'anonymisation automatique du corpus oral Repérer et étudier des éléments dans le discours permettant l'identification de la personne par un éventuel utilisateur du corpus.
Méthodologie	Réflexion préalable sur la nature des données permettant l'identification Définition d'une nouvelle notion : *le faisceau d'indices* qui dépasse celle des entités nommées**.** Développement de la méthode symbolique consistant dans la création de règles d'extraction (patrons) sous forme de graphes en utilisant l'outil de l'annotation des entités nommées CasSys et en l'adaptant au corpus traité et à l'objectif visé.
Données traitées	Transcriptions des enregistrements en face-à-face d'ESLO1
Résultats	Corpus : un échantillon du corpus ESLO1 (112 entretiens en face-à-face) annotés en entités nommées et indices d'identification Elaboration de la typologie des éléments adéquats à notre besoin, à partir de la typologie de la campagne d'évaluation Ester2 Construction d'une grammaire locale (une série de graphes) permettant la reconnaissance et le balisage automatiques des informations personnelles sur le locuteur Analyse de ces informations dans le corpus
Difficultés rencontrées Contraintes et limites	Le processus d'identification peut être direct ou indirect. Ce dernier est le plus sovent utilisé dans le discours. C'est effectivement le recoupement de plusieurs indices qui permet de lever l'identité de la personne. De plus, toute information personnelle n'est pas identifiante. Repérer tous ces indices automatiquement ne suffit pas, il est nécessaire de procéder au filtrage manuel de données pour décider si tel ou tel indice doit être anonymisé. Le processus d'anonymisation ne s'arrête pas à la reconnaissance des entités nommées classiques, car il s'agit aussi du repérage des éléments d'identification hors noms propres. D'une part la reconnaissance des entités nommées ne suffit pas à repérer toutes les informations concernant le sujet parlant et, d'autre part, toutes les entités nommées ne doivent pas être anonymisés. L'entité nommée repérée doit être étiquetée selon son rapport avec le locuteur. C'est dans ce contexte du lien avec la personne qu'elle devient identifiante. Des connaissances extralinguistiques nécessaires pour repérer certains indices ne peuvent pas être totalement modélisables. Dans les entretiens on trouve aussi des informations occasionnelles « imprévisibles », et donc non homogènes, sur le locuteur. Ces informations sont difficilement repérables automatiquement.

	La présence de multiples disfluences (hésitations, répétitions, reformulations, amorces etc.) qui peuvent intervenir à différents moments dans le discours comme dans *je m'appelle euh Patrick Mallon* rend cette tâche difficile mais réalisable. Les informations de nature personnelle varient d'une manière non homogène dans le corpus et on ne peut jamais atteindre une liste exhaustive de toutes les reformulations possibles. Non prise en compte des informations prosodiques
Originalité et apport du travail	Le travail sur l'anonymisation du corpus ESLO a permis : - de constater que toutes les informations personnelles n'identifient pas la personne mais qu'en revanche une combinaison de certaines d'entre elles constituent un faisceau qui dans un certain contexte, le plus souvent extralinguistique, contribuent à l'identification. - définir et décrire les éléments permettant l'identification du locuteur. Le traitement automatique de ces éléments ne peut pas se satisfaire du repérage des entités nommées. D'une part, ils dépassent par leur diversité les entités nommées et, d'autre part, les entités nommées repérées doivent fournir des informations sur le locuteur. - de développer la typologie des indices d'identification que nous avons utilisée pour leur annotation. - de montrer les difficultés du traitement automatique de l'anonymisation du corpus oral : les informations occasionnelles, imprévisibles et non homogènes, la prise en compte du contexte extralinguistique et la variation infinie des reformulations possibles à l'oral. - d'étudier la subjectivité dans le discours sous un nouvel angle : à partir de l'étude du fasceau d'indices d'identification. Le travail sur l'anonymisation a été effectué sur le corpus oral.

3.4.6.2. *Perspectives*

Les outils du TAL sont utiles et efficaces pour le repérage des informations personnelles. Cette tâche réussit mieux dans les corpus écrits où ces informations sont présentées d'une manière plus au moins homogènes. La tâche se complique pour le discours oral. Si l'on réussit efficacement à repérer automatiquement les éléments personnels, il est nécessaire ensuite de choisir parmi ces éléments ceux qui permettent l'identification de la personne. Ce filtrage ne peut se faire que manuellement à l'heure actuelle. La perspective d'une automatisation de cette détection est un défi que les recherches en TAL pourraient relever.

Pour aider la validation manuelle, la distinction pourrait se faire entre les éléments les plus sensibles à l'anonymisation, c'est-à-dire ceux qui apportent une information plus importante et plus spécifique, et ceux qui sont plus généraux. Ainsi, pour distinguer entre les noms de famille rares comme *Eshkol* ou *Kanaan* et très répandues *Dupond* ou *Durand*, on pourrait s'appuyer, dans le cas du corpus des ESLO, sur une information concernant la fréquence d'un nom propre, éventuellement pondérée par des critères géographiques. De la même manière, le locuteur peut désigner son métier par un seul mot *enseignant* ou en précisant *professeur de physique.* Ce passage d'un seul nom à un groupe nominal plus étendu grâce aux modifieurs concerne n'importe quelle caractéristique. On pourrait ainsi attribuer plus de poids à ces éléments sensibles à l'anonymisation ce qui diminuerait le nombre des indices candidats à l'anonymisation et de cette manière aiderait la validation manuelle.

Dans le faisceau d'indices, l'étude des éléments ne faisant pas partie des entités nommées comme actions, événements, activités sociales, doit être approfondie d'autant plus qu'elle permet d'apporter une information sur le profil sociologique du locuteur. Le module développé tient compte de ces indices mais leur liste n'est pas exhaustive. Pour un travail futur, j'envisage d'étudier avec précision dans le corpus ESLO, tous les éléments anonymisés par une procédure manuelle afin d'affiner la typologie du faisceau d'indices.

Enfin, je souhaite approfondir mon investigation sur les entités nommées en prenant en compte la subjectivité. La désignation d'une personne ou d'un lieu est un processus social réapproprié subjectivement. En désignant quelqu'un par son prénom ou par son statut, on ajoute une information sur ses origines ou son statut. L'univers subjectif du locuteur se donne à lire dans cette référence à une personne ou à un lieu donnés.

3.5. Repérage, annotation et analyse des reformulations paraphrastiques

Le troisième type d'annotation du corpus oral sur lequel j'ai travaillé porte sur un autre phénomène très fréquent à l'oral : les reformulations paraphrastiques. Il s'agit de repérer automatiquement et d'annoter ces paraphrases dans ESLO. Le travail a été effectué en collaboration avec Natalia Grabar (STL, Université de Lille) et il se poursuit actuellement.

3.5.1. Résumé du travail

Le phénomène de la paraphrase est bien étudié dans le domaine de la linguistique et du TAL. Nous nous sommes intéressées à la reformulation paraphrastique, si fréquente dans le discours oral où le locuteur essaie de reformuler son énoncé pour différentes raisons : explication, précision, correction etc. Ce phénomène peut être considéré comme faisant partie des disfluences de l'oral car il s'agit d'une interruption du flux de parole.

Les méthodes de détection automatique de paraphrase sont souvent fondées sur les propriétés paradigmatiques des éléments linguistiques paraphrasés et la possibilité d'une substitution. Pour acquérir automatiquement les entités linguistiques paraphrastiques, les chercheurs exploitent des corpus comparables (décrivant le même événement, par exemple), des corpus bilingues ou monolingues parallèles et des corpus monolingues où les entités linguistiques (mots, phrases) ayant la même distribution dans le corpus sont considérées comme de bons candidats pour le processus. Ces travaux se font à partir de corpus écrits.

Notre démarche est différente. Nous avons proposé une méthode de détection des RP (reformulations paraphrastiques) dans un corpus oral monolingue. Notre approche est syntagmatique, par différence avec les approches souvent utilisées en TAL pour effectuer cette tâche. Nous travaillons sur les paraphrases dans le corpus oral où elles sont produites d'une manière naturelle et spontanée sans être le résultat d'un alignement ou d'une traduction. La méthode tient compte également de la nature des données traitées. Cela concerne la reconstitution d'énoncés sans marque typographique dans les transcriptions, la prise en compte des disfluences et l'utilisation d'outils adaptés à l'oral. Nous utilisons une définition plus large de la paraphrase qui ne se limite pas à l'équivalence sémantique entre deux segments. Nous considérons ainsi que la paraphrase peut être utilisée non seulement pour reformuler une idée mais aussi pour la décrire, l'exemplifier, la préciser ou l'expliquer.

Nous avons mis en œuvre un processus de détection automatique. Ce processus permet de désambiguïser automatiquement les emplois paraphrastiques et non paraphrastiques.

Les 611 tours de parole dans 59 entretiens d'ESLO1 et 498 tours de parole dans 37 entretiens d'ESLO2 ont été annotés multidimensionnellement en reformulations paraphrastiques. Le jeu d'étiquettes établi tient compte de différents aspects de modifications linguistiques. En se fondant sur ce corpus annoté, une étude des liens formels entre les deux segments paraphrasés et une étude des emplois de trois marqueurs *c'est-à-dire, je veux dire* et *disons* ont été réalisées. L'analyse du corpus annoté a montré des particularités des reformulations paraphrastiques qui n'ont pas été relevées dans les travaux antérieurs, à savoir, par exemple, la paucité des liens formels au niveau morphologique, syntaxique et même lexicale entre les deux segments paraphrasés.

Les travaux sur les reformulations paraphrastiques dans le corpus oral sont décrits dans (Eshkol-Taravella et Grabar 2014a,b, à paraître, Grabar et Eshkol-Taravella 2015).

3.5.2. Etat de l'art

La notion de la paraphrase a fait l'objet de nombreuses études linguistiques. Je distinguerai deux approches selon que :

- la paraphrase est étudiée du point de vue contextuel (Culioli 1976, Fløttum 1995, Fuchs 1994, Martin 1976, Roulet 1987, Rossari 1990, Vezin 1976) ;
- elle est définie à travers les transformations linguistiques aux différents niveaux (morphologique, lexical, syntaxique, sémantique) (Bhagat et Hovy 2013, Melčuk 1988, Vila *et al.* 2011).

Cette notion est bien représentée en TAL (Androutsopoulos et Malakasiotis 2010, Bouamour 2012, Ibrahim *et al.* 2003, Madnani et Dorr 2010, Pasça et Dienes 2005, Quirk *et al.* 2004 etc.). Les méthodes utilisées sont souvent fondées sur les propriétés paradigmatiques des éléments linguistiques paraphrasés et la possibilité d'une substitution. Pour acquérir automatiquement les entités linguistiques paraphrastiques, les chercheurs exploitent des corpus comparables (décrivant le même événement, par exemple), des corpus bilingues ou monolingues parallèles et des corpus monolingues où les entités linguistiques (mots, phrases) ayant la même distribution dans le corpus sont considérées comme de bons candidats pour le processus. La méthodologie repose sur l'utilisation d'un ou plusieurs corpus comparables ou parallèles afin de « se procurer » des paraphrases. Les travaux décrits dans cette partie se font sur des corpus écrits.

Le critère commun qui définit la paraphrase et sur lequel une majorité de chercheurs semblent s'accorder concerne une équivalence sémantique entre les expressions linguistiques en relation de paraphrase. Un autre critère est la présupposition du lien formel obligatoire entre les deux segments paraphrasés aux différents niveaux linguistiques (morphologique, lexical, syntaxique etc.) qui permet d'une part de modéliser cette relation et d'autre part de la détecter automatiquement.

Notre démarche est différente. Tout d'abord, nous acceptons une définition plus large de la paraphrase. Nous considérons que la paraphrase peut être utilisée non seulement pour reformuler une idée mais aussi pour la décrire, l'exemplifier, la préciser ou l'expliquer. La méthode que nous utilisons est inspirée par l'exemple des corpus oraux dans lesquels les paraphrases sont produites d'une manière naturelle et spontanée sans être un produit de l'alignement ou de la traduction. Nous avons mis en œuvre un processus de détection automatique de cette relation fondée sur une approche syntagmatique et non paradigmatique utilisée souvent dans ce cas en TAL.

Nous nous sommes intéressées dans notre travail à un phénomène qui déborde la paraphrase *stricto sensu*. Il s'agit des reformulations paraphrastiques. De manière générale, il est considéré que la reformulation est une activité du locuteur qui s'appuie sur un segment déjà produit dans son propre discours ou dans celui de son interlocuteur, avec ou sans l'emploi d'un marqueur (Gülich et Kotschi 1987, Kanaan 2011). Tout acte de reformulation dans le discours oral n'introduit pas une paraphrase (Rossari 1990). De ce point de vue, on distingue deux catégories de marqueurs : les marqueurs de reformulation non-paraphrastique (*e.g. en somme, en tout cas, de toute façon, enfin* etc.) et les marqueurs de reformulation paraphrastique (ou MRP), comme *c'est-à-dire, autrement dit, je m'explique, ça veut dire, en d'autres termes* (Gülich et Kotschi 1983, Rossari 1990, 1994). Les auteurs distinguent ceux qui ont pour tâche principale d'établir une relation paraphrastique (*e.g. c'est-à-dire, autrement dit*) et ceux qui ne montrent ce rôle que dans des contextes précis. En outre, les propriétés sémantiques des MRP permettent d'instaurer une relation de paraphrase même entre des segments qui n'entretiennent aucune équivalence sémantique constatable par ailleurs (Rossari 1994).

Le travail présenté dans cette partie a été effectué sur 260 entretiens d'ESLO1 totalisant 2 349 829 occurrences de mots et 308 entretiens d'ESLO2 totalisant 1 412 891 occurrences de mots. Le travail a porté sur trois MRP : *c'est-à-dire, je veux dire* et *disons* formés à partir du même verbe *dire*.

3.5.3. Méthodologie

La méthologie adoptée est visualisée dans la Figure 3.

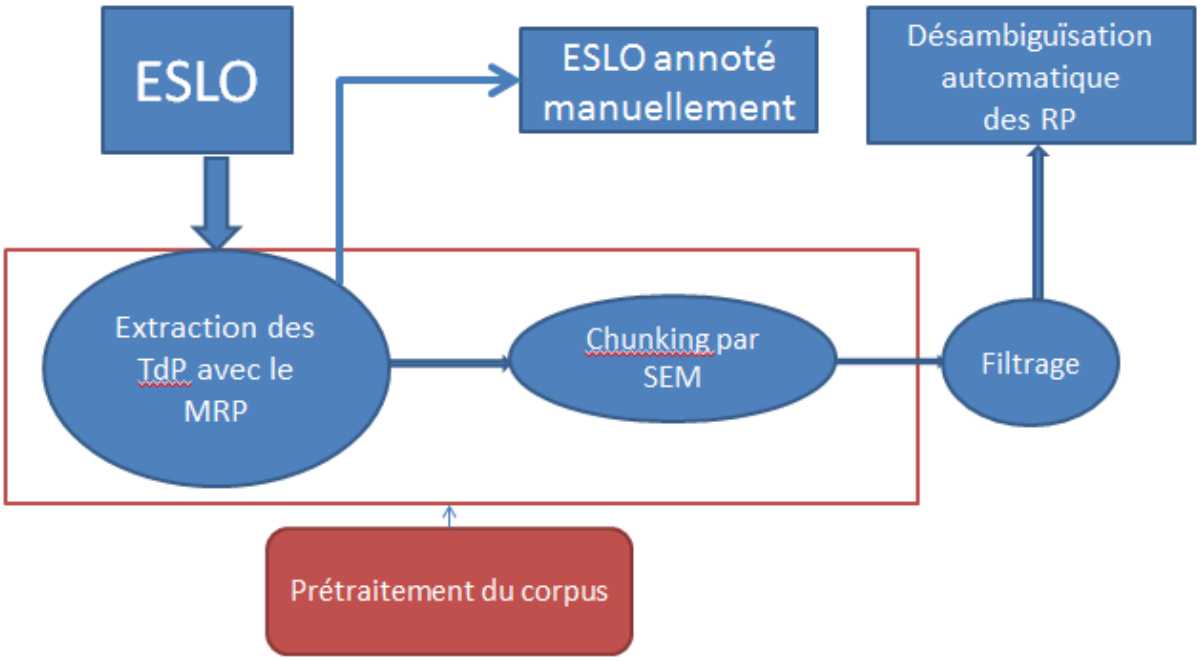

Figure 3 : Méthodologie

260 entretiens d'ESLO1 et 308 entretiens d'ESLO2 transcrits ont été prétraités. Tout d'abord, les fichiers ont été segmentés en tours de parole dans le but de reconstituer les énoncés : l'énoncé commence avec le changement de locuteur et, en cas de chevauchement, les segments correspondants sont associés aux énoncés de chacun des locuteurs impliqués : lorsqu'un locuteur continue de parler après un chevauchement, son tour de parole est prolongé d'autant. On extrait ensuite les énoncés contenant l'un des marqueurs étudiés. On obtient 476 tours de parole dans 54 entretiens d'ESLO1 et 394 tours de parole dans 30 entretiens d'ESLO2. L'annotation manuelle est effectuée ensuite sur ce corpus selon la méthodologie suivante : les deux annotateurs annotent d'abord le même corpus séparement, ensuite ils annotent le même corpus ensemble pour arriver à un consensus en cas de discordance. On obtient ainsi trois versions d'annotation du même corpus, la dernière étant considérée comme définitive. Les tours de parole établis sont traités par SEM appris sur l'échantillon d'ESLO (Tellier *et al.* 2013, 2014) qui les étiquette en POS et les annote en chunks. Cette sortie en chunks est utilisée par Natalia dans le processus de désambiguïsation automatique pour distinguer emplois paraphrastiques et non paraphrastiques. Pour cette tâche, elle applique plusieurs filtres qui vérifient le contexte d'emploi du marqueur.

3.5.4. Annotation des reformulations paraphrastiques

3.5.4.1. Objectifs

L'annotation manuelle répond à plusieurs objectifs. Tout d'abord, elle permet de faire une première distinction entre l'emploi paraphrastique et non paraphrastique de la reformulation. Le corpus annoté manuellement est un corpus de référence par excellence qui permet de procéder à l'analyse du phénomène annoté. Il peut servir aussi de modèle pour l'apprentissage automatique et permettre l'évaluation du module de détection automatique du phénomène (Grabar et Eshkol-Taravella 2015). Le travail sur l'annotation manuelle a permis d'émettre un

certain nombre d'hypothèses qui ont été prises en compte par le module de détection automatique.

L'annotation proposée est multidimensionnelle, inspirée par certains travaux sur la reformulation. Pour évaluer l'annotation manuelle, l'accord inter-annotateur a été calculé en utilisant le kappa de Cohen.

3.5.4.2. Convention et format de l'annotation

L'annotation porte d'une part sur les deux segments en relation de paraphrases, mais aussi sur la relation établie par le MRP de manière générale. Elle est effectuée sous forme de balises : l'information contenue dans les balises désigne les catégories syntaxiques des segments et les attributs indiquent les informations sur le processus et le fonctionnement de la paraphrase établie grâce au MRP dans chaque cas. Le corpus a été annoté selon les conventions prédéfinies (Annexe 12). Les informations annotées sont de nature

Syntaxique :

- catégorie syntaxique (N, A, V, Prep…) ou type de constituant syntaxique (NP, VP, AP, PP).

Cela permet de vérifier une éventuelle équivalence syntaxique entre les deux segments.

- modifications syntaxiques : passif/actif… ;

Lexicale :

- relations lexicales : hyperonyme, hyponyme, synonyme, antonyme, instance, méronyme ;
- modifications lexicales : remplacement, suppression, ajout ;

Morphologique :

- modifications morphologiques : flexion, dérivation, composition ;

Pragmatique :

- relations pragmatiques : définition, explication, exemplification, précision, dénomination, résultat, correction linguistique, correction référentielle, équivalence.

Cette relation reprend des typologies proposées dans la littérature (Gülich et Kotschi 1987, Kanaan 2011).

Observons quelques exemples. Dans le cas suivant :

> *pendant nous avons fait grève à la Régie Renault euh de <NP1>**Saint Jean de la Ruelle**</NP1><MRP>**c'est-à-dire**</MRP> <NP2 rel-lex= »mer(Saint Jean de la Ruelle/Orléans) » rel-pragm= »cor-ref »>**Orléans**</NP2> parce que c'est ça fait partie d'Orléans* [ESLO1_ENT_149_C]

le lieu *Saint Jean de la Ruelle* est remplacé par *Orléans* en utilisant le MRP *c'est-à-dire.* Il s'agit d'un cas de méronymie et de correction référentielle déclenchée par une supposée incompréhension de l'interlocuteur. Ce type de correction est surtout attesté dans ESLO1 où l'interviewer était un Anglais. Dans un autre exemple,

> *on fait ce que l' on appelle <NP1>**un carton**</NP1> <MDR>**c'est-à-dire**</MDR> le le <NP2 rel-lex= »hyper(carton/dessin) » modif-lex= »remplacement(un/ce…-là) remplacement(carton/dessin) ajout(Adj=agrandi) » rel-pragm= »prec »>**ce dessin-là agrandi**</NP2> mais à la grandeur de la fenêtre* [ESLO1_ENT_002_C]

le groupe nominal *un carton* est paraphrasé par *ce dessin-là agrandi.* Plusieurs modifications ont été mises en œuvre pour effectuer cette correction afin de préciser l'information fournie. Le locuteur a remplacé *carton* par son hyperonyme *dessin,* le déterminant indéfini *un* est devenu un démonstratif avec le composé discontinu *ce…-là* ; il a ajouté également un adjectif *agrandi* qualifiant le substantif. Enfin, dans

> *euh <VP1>**démocratiser l'enseignement**</VP1> <MRP>**c'est-à-dire** </MRP> <VP2 rel-lex= »syn(démocratiser/permettre à tout le monde) syn(enseignement/faculté) » modif-lex= »ajout(rentrer à) » rel-pragm= »explic »>**permettre à tout le monde de rentrer en faculté**</VP2>* [ESLO1_ENT_121_C]

c'est un groupe verbal *démocratiser l'enseignement* qui est paraphrasé par *permettre à tout le monde de rentrer en faculté.* Le locuteur explique ce qu'il entend par *démocratiser l'enseignement.* Pour cela, il remplace le verbe *démocratiser* par son synonyme *permettre à tout le monde* et *enseignement* par *faculté.*

3.5.4.3. Résultats et Discussion

L'annotation manuelle a permis de réfléchir sur les emplois de trois marqueurs en discours. D'une manière générale, *c'est-à-dire* est le plus fréquent. *C'est-à-dire* et *disons* représentent 30% de cas dans ESLO1 et 16% dans ESLO2 en tant que MRP. Le marqueur *je veux dire* fonctionne comme un MRP entre les deux segments dans 14% (ESLO1) et 19% (ESLO2) des cas. Il est donc moins utilisé dans ESLO1 et plus fréquent que les deux autres marqueurs dans ESLO2. Faut-il y voir une évolution diachronique de la langue telle que ce marqueur reprendrait la fonction de MRP des deux autres marqueurs. Une autre explication est liée au contexte d'enregistrement d'ESLO2 qui favorise mieux la parole spontanée et incite l'interviewé à parler d'une manière moins formelle. La structure du marqueur *je veux dire* (le pronom personnel à la première personne *je* et le verbe modal *vouloir*) présuppose une implication plus forte de l'énonciateur symptomatique d'un changement en cours.

L'analyse du corpus annoté a permis de vérifier quelques hypothèses émises sur la paraphrase et la reformulation paraphrastique par (Gülich & Kotschi 1983; Rossari 1994) qui constatent le parallélisme syntaxique entre l'entité source et l'entité paraphrasée. Cette affirmation est vérifiée par l'annotation car dans la majorité des cas (60 %), il existe une équivalence syntaxique entre les éléments en relation de paraphrase. On pourrait expliquer ce phénomène aussi par l'aspect subjectif de l'annotation manuelle. La décision sur la délimitation des frontières des segments paraphrasés revient aux annotateurs qui peuvent être « influencés par une catégorie syntaxique du premier segment. Les études sur les paraphrases citées en 3.5.2 convergent vers l'existence du lien formel entre les segments paraphrasés. Ce lien est modélisé à travers les transformations morphologiques, lexicales, syntaxiques. Une annotation multidimensionnelle tenant compte de ces aspects a fait la démonstration du contraire. Les modifications morphologiques ne représentent que 10% des reformulations annotées. Nous n'avons observé qu'un seul exemple de modification syntaxique (actif/passif) dans le corpus annoté ESLO1. Les modifications faites au niveau lexical représentent 57% de toutes les reformulations annotées dans ESLO1 et 30% dans ESLO2. En ce qui concerne les relations lexicales, elles occupent 75% des cas annotés dans ESLO1 et 58% dans ESLO2. Du point de vue diachronique, nous observons donc un recul des liens lexicaux entre les éléments contenus dans les deux segments paraphrasés. En conclusion, nous pouvons constater, à l'intérieur du corpus oral analysé, qu'il existe très peu de repères formels pour détecter des segments en relation de paraphrase dans ce type de constructions. Cette remarque est valable

aux niveaux morphologique, lexical et syntaxique. Cette tendance semble se confirmer dans ESLO2 ce qui pourrait s'expliquer par les conditions d'enregistrement.

L'annotation des relations pragmatiques a permis de distinguer trois processus fonctionnels de la RP : le locuteur (i) ajoute une nouvelle information (explication, précision, exemplification, justification et définition), (ii) répète la même information mais d'une autre façon, dans ce cas il est possible de supprimer le MRP et de changer les segments de place sans modifier le sens de l'énoncé (paraphrase) et (iii) synthétise (résultat) ou dénomme (dénomination) ce qui vient d'être dit. Ce constat est confirmé par l'association entre une relation pragmatique annotée et la taille des segments calculée en nombre de mots. Nous distinguons trois cas :

- le segment 2 est plus long que le segment 1 : le locuteur précise, définit, explique ou exemplifie ses propos ;

- le segment 1 est plus long que le segment 2 : le locuteur conclut, raccourcit ce qui a été dit (*res*) ou il donne le nom à ce qui a été annoncé précédemment (*dénom*) ;

- les deux segments sont équivalents : il s'agit de la paraphrase « pure » ou de la correction linguistique.

Ce lien entre la fonction pragmatique de la reformulation et la différence de longueur entre les segments paraphrasés pourrait être mobilisé comme critère de classification automatique de la reformulation.

D'autres observations moins significatives peuvent être ajoutées. Parmi les catégories syntaxiques, les plus fréquentes sont les propositions suivies des groupes nominaux. Dans un corpus oral, reformuler des énoncés entiers semble « naturel ». La relation lexicale la plus souvent annotée est la synonymie, ce qui semble aussi logique. Le lien de méronymie entre les éléments des deux segments augmente dans ESLO2. Nous avons considéré comme méronymie les cas du rapport /*partie vs tout*/ en ajoutant les liens par association. La conduite de l'entretien dans ESLO2 pouvait favoriser la présence de ces liens associatifs entre les éléments. En ce qui concerne les modifications lexicales, le locuteur paraît préférer à l'ajout de nouveaux mots ou à la suppression du premier segment la substitution d'un mot ou d'un sous-segment.

3.5.5. Règles de désambiguïsation automatique

L'annotation manuelle des RP a permis de définir des règles de désambiguïsation automatique. Le traitement automatique consiste alors à décider si, autour d'un MRP, il existe une relation de reformulation paraphrastique ou non.

Parmi les trois marqueurs étudiés, *disons* est le plus polysémique. Les trois contextes où il n'indique pas de reformulation paraphrastique sont :

- lorsque *disons* suit le pronom personnel *nous.*

Dans ce cas, il s'agit d'un verbe *dire* conjugué au présent de l'indicatif ou employé à l'impératif ;

- l'insertion de *disons* dans des suites argumentatives (*e.g. par contre, mais, en revanche, au contraire, cependant*).

Dans ce cas, il s'agit du verbe *dire* et de l'introduction d'une nouvelle information par opposition à ce qui a annoncé précédemment[36].

- lorsque *disons* se retrouve au milieu d'un syntagme syntaxique.

[36] Ce contexte pourrait concerner également les deux autres marqueurs *c'est-à-dire* et *je veux dire* mais je n'en ai pas trouvé d'attestations dans ESLO.

Dans ces cas, il suspend le déroulement syntagmatique sans rien ajouter à la sémantique de l'énoncé et peut être considéré comme une disfluence. Les syntagmes interrompus par *disons* ont des constructions diverses :

- V disons GPREP : verbe transitif indirect suivi de son complément

  une personne ***habitant disons à cinq kilomètres de de chez vous***

- V disons GN : verbe transitif suivi de son complément

  je ***vocalise disons une demi-heure***

- GN disons PREP GN / GN disons PrREL V: groupe nominal étendu

  la porte *euh* ***disons de cette pièce***

  les jeunes disons qui se marient

- GN disons ADJ : substantif suivi de l'adjectif épithète qui le qualifie

  c' est ***un chant disons assez vulgaire***

  un ancien appartement euh disons bourgeois

- V disons par GN : construction au passif

  on n' a pas ***été suivi disons par les parents***

- VAUX disons PP : verbe conjugué au temps composé

  comment vous ***avez disons choisi***

Deux règles générales concernant trois marqueurs ont été fixées :

- si le MRP est placé en début ou en fin d'énoncé, le contexte n'est pas jugé suffisant pour établir une paraphrase ;
- dans le cas où les MRP sont entourés des marqueurs discursifs (*donc, enfin, quoi…*), *euh d'hésitation*, interjections (*ben, mh, ouais*), amorces (*s-*) etc. répétés, nous considérons qu'ils font partie des disfluences de l'oral.

3.5.6. Bilan

3.5.6.1. Compte rendu

Objectif visé	Annoter les reformulations paraphrastiques autour de trois marqueurs *c'est-à-dire, je veux dire* et *disons* dans le corpus oral. Repérer automatiquement les tours de parole contenant la relation de paraphrase En se fondant sur l'annotation : - étudier ce phénomène du point de vue diachronique, - étudier les emplois de ces trois marqueurs dans le discours - vérifier et quantifier les liens formels entre les deux segments paraphrasés
Méthodologie	Méthode à base de règles pour distinguer, au sein d'un ensemble de tours de parole comportant un des trois marqueurs étudiés (*disons, c'est-à-dire, je veux dire*), les tours de parole qui comportent des reformulations paraphrastiques.
Données traitées	Transcriptions des enregistrements face-face d'ESLO1 et d'ESLO2
Difficultés rencontrées	Deux caractéristiques de l'oral : la présence des disfluences et le manque de ponctuation dans les fichiers de transcription. Ambiguïté de certains marqueurs traités. Prise de décision « objective » sur certains paramètres annotés (relations pragmatiques ou détermination des frontières des segments paraphrasés).
Résultats	Corpus : 612 tours de parole dans 260 entretiens d'ESLO1 et 500 tours de parole dans 308 entretiens d'ESLO2 annotés multidimensionnellement en reformulations paraphrastiques Etude des liens formels entre les deux segments paraphrasés Etude (synchronique et diachronique) des emplois de trois marqueurs : *c'est-à-dire, je veux dire* et *disons*
Contraintes et limites	Non prise en compte des informations prosodiques Le traitement se fait au sein d'un tour de parole La structure étudiée est *segment1 MRP segment2*, en sorte que les reformulations sans marqueurs ou avec des marqueurs distants, ne sont pas prises en compte.
Originalité et apport du travail	Méthode de détection des RP dans un corpus oral monolingue. Approche syntagmatique différentes des approches utilisées d'ordinaire en TAL pour effectuer cette tâche. Méthode tenant compte de la nature des données traitées. Utilisation d'une définition plus large de la paraphrase qui ne se limite pas à l'équivalence sémantique entre deux segments. Vérification des hypothèses émises dans des travaux antérieurs sur la paraphrase et la RP : confirmation du paralléllisme syntaxique entre les deux segments paraphrasés et constat de la rareté des liens formels au niveau morphologique, lexical et syntaxique entre les deux segments paraphrasés.

	Constat du lien entre la fonction pragmatique et la taille des segments paraphrasés Quelques observations sur les emplois de trois marqueurs du point de vue diachronique

3.5.6.2. Perspectives et travaux en cours

Les travaux sur la reformulation paraphrastique se poursuivent. Une méthode automatique fondée sur l'apprentissage avec les CRF à partir du corpus annoté manuellement afin de détecter les segments paraphrasés a été mise en œuvre (Grabar et Eshkol-Taravella 2015).

J'ai aussi continué l'annotation manuelle selon les conventions définies en collaboration avec les étudiants dans le cadre du cours « Enrichir le corpus » en ajoutant d'autres marqueurs : *ça veut dire, j'allais dire, notamment, autrement dit, en d'autres termes, en d'autres mots*. Deux corpus supplémentaires ont été annotés selon les mêmes conventions : le corpus médiatique Le Monde extrait du web durant un mois et le corpus du forum de discussion Doctissimo. La stratégie d'annotation suivante a été proposée aux étudiants. Les deux étudiants annotent d'abord le même corpus séparement, ensuite ils annotent le même corpus ensemble pour arriver à un consensus en cas de discordance. Cette stratégie présente de nombreux avantages. D'abord, on obtient trois versions d'annotation du même corpus et la dernière est considérée comme définitive, ce qui est d'un grand intérêt pour des études comparatives. Ensuite, les étudiants sont obligés de parvenir à un consensus et d'inventer des critères bien définis pour les cas difficiles. Enfin, ils restituent une analyse des difficultés rencontrées et des critères qui ont permis l'accord.

Plusieurs perspectives s'offrent pour ce travail. Tout d'abord, nous voudrions tester les modèles développés pour la détection automatique des tours de parole avec la reformulation paraphrastique sur d'autres corpus. Ainsi, pourrons-nous étudier si des reformulations paraphrastiques introduites par différents marqueurs présentent des régularités similaires. D'autres perspectives consistent à appréhender ces données selon d'autres points de vue. Par exemple, nous pouvons prendre en compte et analyser conjointement les éléments prosodiques et acoustiques associés aux différents tours de parole. Notre hypothèse est que les tours de parole avec les reformulations paraphrastiques montrent des différences par rapport aux tours de parole avec les marqueurs étudiés mais n'introduisant pas de reformulations paraphrastiques. De cette manière, le filtrage entre ces deux types de tours de parole peut reposer également sur ce critère. En possession de corpus annotés de nature différente, nous nous proposons de procéder à une analyse comparative. Le traitement des relations de paraphrase qui est effectué au sein d'un même tour de parole pourrait être élargi à plusieurs tours de parole. Nous voudrions enfin étudier l'emploi des reformulations paraphrastiques en croisant les annotations avec les critères sociologiques des locuteurs.

3.6. Des recettes d'omelettes

C'est le dernier travail présenté dans ce mémoire portant sur l'oral. Il concerne l'annotation d'informations de nature sémantique et pragmatique, ce qui le rapporche de ce qui précède. Les travaux décrits jusqu'ici sont des travaux qui relèvent du TAL. J'ai pu participer au développement d'« instruments » pour le traitement des corpus, qu'il s'agisse de créer un analyseur syntaxique ou un annotateur en informations personnelles sur la personne, de détecter des reformulations paraphrastiques dans les entretiens d'ESLO. Le travail que je présente dans cette partie concerne, contrairement aux autres travaux, le domaine de la linguistique outillée.

Je me suis intéressée au sous-corpus ESLO1 composé des réponses à la question *Comment faites-vous une omelette ?* La recette d'un plat aussi banal que l'omelette est un exemple, par excellence, d'un discours « neutre », sans connotation subjective. La recherche a montré pourtant que l'objet si commun peut être personnalisé dans le discours oral.

3.6.1. Résumé du travail

Les recettes de cuisine font partie des textes procéduraux et sont au centre de recherches dans différents domaines : linguistique, TAL, représentation des connaissances. Si les recettes écrites ont fait l'objet de différentes études, il n'en va pas de même avec des recettes transmises oralement.

Nous émettons l'hypothèse que les recettes proposées dans le cadre de l'interview doivent contenir en plus des informations classiques propres aux recettes, d'autres types d'informations appartenant à la communication orale comme les marques de l'énonciation, les éléments phatiques, la diversité des niveaux du discours (interpellations de l'auditeur, anecdotes etc.), les connecteurs ou les disfluences.

Ce travail s'inscrit dans le domaine de la linguistique du corpus. L'étude des recettes est fondée sur des outils et des méthodes informatiques existants : concordancier, feuille de style XSLT, tableur Excel. L'annotation manuelle proposée tient compte des spécificités du corpus et a permis, d'une part, de typer et de cette manière de modéliser les informations contenues dans le corpus et d'autre part, d'extraire automatiquement ces informations afin de pouvoir les analyser quantitativement et qualitativement.

Une observation manuelle du corpus a permis de distinguer quatre types d'informations :
- *les actions élémentaires* correspondant aux six étapes rudimentaires de la recette typique de l'omelette ;
- *les expansions* concernant les éléments secondaires liés à la recette ;
- *les développements latéraux* regroupant des données qui entretiennent un rapport indirect avec la recette ;
- *les méta-commentaires* concernant la situation de communication.

L'analyse des trois recettes écrites prises en référence a permis de distinguer six opérations fondamentales nécessaires à la préparation d'omelette : *casser les œufs, les battre, saler (poivrer), chauffer dans la poêle la matière grasse, verser les œufs battus, (faire) cuire.* Toutes ces informations ont été annotées et extraites du corpus pour l'analyse quantitative et qualitative, leurs variations et leur ordre d'enchaînement ont été examinées.

L'étude du corpus de recettes d'un plat aussi simple qu'une omelette a révélé des phénomènes et variations linguistiques très intéressants. La différence du corpus étudié par rapport aux corpus classiques des recettes de cuisine tient à son mode oral de transmission et de restitution. Une autre différence tient au fait qu'il s'agit d'une réponse à une question posée dans un cadre inattendu, celui d'un dialogue où la présence de l'interlocuteur doit être prise en compte dans l'analyse.

Le travail sur le corpus des recettes d'omelettes a été decrit dans (Bergounioux et Eshkol, Eshkol et Bergounioux, à paraître)

3.6.2. Etat de l'art

Les recettes de cuisine sont au centre de recherches dans différents domaines.

En premier lieu, elles font partie de ce qu'on appelle les textes procéduraux. Selon (Heurley 1997), le texte procédural se définit par sa fonction principale. « La notion de texte procédural […] fait référence à tous les textes dont la fonction principale est de communiquer des procédures en vue d'une exécution ponctuelle […] ou d'un apprentissage à long terme censé permettre l'acquisition d'un savoir-faire nouveau dans un domaine particulier […]. Les textes procéduraux ont donc avant tout une fonction pragmatique […]. Cette catégorie regroupe, notamment, les modes d'emploi, les notices explicatives, les manuels et les guides d'utilisation, les consignes de sécurité, les recettes de cuisine, les do-listes utilisées en aéronautique etc. […] ». L'auteur affirme en 2001 que les informations procédurales « spécifient, quant à elles, des opérations ou des actions que l'utilisateur doit exécuter (instructions positives), ou au contraire, s'abstenir d'exécuter (instructions négatives). » (Heurley 2001 : 30-31). (Adam 2001) montre qu'un texte procédural (comme une recette de cuisine) ne peut être assimilé à un récit. Il s'agit de situations spécifiques orientées ayant une finalité pratique. En conséquence, ces textes, aussi divers soient-ils, comportent un certain nombre de marqueurs linguistiques de la position illocutoire de l'auteur du texte : verbes d'action, d'ordre et de conseil etc. (Virbel 1997) a proposé une caractérisation des textes procéduraux (qu'il appelle *textes de consignes*) fondée sur la théorie des actes de langage. La recette constitue aussi un « genre » de texte selon (Rastier 2001).

Les recettes de cuisine actualisent différentes variations qui ont fait l'objet d'études en sciences du langage. A titre d'exemple, je citerai le travail de nature diachronique de (Colson 2012) où l'auteure a étudié la recette du blanc-manger au travers de différents livres de cuisine des XVII^{e} et XVIII^{e} siècles en analysant en particulier des caractéristiques de forme et de contenu. Les variantes « lexicales, morphologiques, syntaxiques » et « leurs différents effets sur la conception et la réception du texte » ont été observées.

Dans le domaine du TAL, les corpus contenant des recettes de cuisine sont au centre des campagnes d'évaluation comme la compétition internationale Computer Cooking contest (CCC) ou le défi DEFT, un atelier annuel d'évaluation francophone en fouille de textes. En 2013, cette campagne organisée dans le cadre de la conférence TALN2013, a porté sur l'analyse automatique des recettes de cuisine en langue française. Deux actions ont été évaluées :

- la classification de documents

  identifier à partir du titre son niveau de difficulté sur une échelle à 4 niveaux : *très facile, facile, moyennement difficile, difficile*.

  identifier à partir du titre et du texte le type de plat préparé : *entrée, plat principal, dessert*.

  apparier le texte d'une recette à son titre.

- l'extraction d'information

  extraire du titre et du texte d'une recette la liste de ses ingrédients.

Les recettes de cuisine sont utilisées aussi dans le domaine de la représentation des connaissances. Ainsi, (Dufour-Lussier *et al.* 2010) ont cherché à classer les ingrédients selon les techniques culinaires qui y sont appliquées en utilisant une analyse syntaxique et

sémantique dynamique afin de construire une représentation formelle de chaque recette. Le projet de recherches portant sur ce type de corpus TAAABLE (Badra *et al.* 2008) a eu pour objet de construire un système de raisonnement à partir de cas pour la recherche et l'adaptation de textes de recettes de cuisine. Il s'agit d'un système informatique destiné à résoudre des problèmes culinaires, développé dans le cadre du Computer Cooking Contest pour pouvoir permettre des requêtes du type *je désire la recette d'un plat de riz aux champignons*. Dans le cadre de ce projet, l'ontologie de cuisine formalisée comprenant 4552 concepts a été produite[37].

Si les recettes écrites ont fait l'objet de différentes études, il n'en va pas de même avec des recettes transmises oralement.

3.6.3. Corpus oral des recettes d'omelettes

Le corpus est un extrait d'ESLO1. Il s'agit de 96 réponses à la question *Comment est-ce qu'on fait une omelette ? Pourriez-vous m'expliquer comment on fait ? / Pouvez-vous me donner la recette de l'omelette ?* posée au cours des entretiens en face-à-face.

Les recettes répondent à un ensemble de caractéristiques : brièveté, division en ingrédients et actions, lexique spécialisé, présence de verbes annonçant un acte illocutoire etc.

Observons la distribution des différentes catégories syntaxiques. Les mots outils (47% de tous les mots) sont aussi bien représentés que les mots signifiants (53%) (Tableau 7).

En ce qui concerne d'autres catégories, on constate l'absence des pronoms possessifs et très peu de pronoms relatifs (0,9%). Pour le temps verbal, c'est évidemment le présent qui domine avec 87,2%. Les autres temps et modes – imparfait (1,7%), futur (1,5%), conditionnel présent (2,5%), subjonctif présent (2,5%), passé composé (2,1%), plus-que-parfait (0,2%), infinitif (1,3%) – sont moins utilisés voire absents.

Les deux catégories (verbes et adverbes) prépondérantes dans le corpus (après les substantifs) ont fait l'objet de deux études en collaboration avec Gabriel Bergounioux (Eshkol et Bergounioux, Bergounioux et Eshkol, à paraître).

Catégorie syntaxique	**Fréquence relative**
Mots significatifs	
substantif	53,2%
verbes	23,2%
adverbes	17,8%
adjectifs	5,8%

Tableau 7 : Distribution des catégories syntaxiques dans le corpus

Que dire de la réponse à une question aussi simple que l'est celle d'une recette d'omelette ? Quelle information est présente régulièrement dans les réponses ? Laquelle serait plus singulière ? Y a-t-il une différence par rapport aux recettes classiques proposées à l'écrit et

[37] http://wikitaaable.loria.fr/index.php/Food_tree

celles formulées à l'oral ? Toutes ces interrogations ont fait l'objet des recherches exposées dans la partie suivante.

3.6.4. Modélisation de l'information présente dans une recette

L'étude du corpus de recettes d'un plat aussi simple qu'une omelette a révélé des phénomènes et variations linguistiques très intéressantes. La différence du corpus étudié par rapport aux corpus classiques des recettes de cuisine tient à son mode oral de transmission et de restitution. Une autre différence tient au fait qu'il s'agit d'une réponse à une question posée dans un cadre inattendu, celui d'un dialogue où la présence de l'interlocuteur doit être prise en compte dans l'analyse.

Pour pouvoir procéder à la modélisation du contenu dans une recette orale, un micro-corpus de recettes classiques, tiré de manuels de cuisine et de sites adaptés du Web (voir l'Annexe 13), a permis de comparer :

- l'ouvrage de Françoise Bernard[38], désormais R1,
- le site www.marmiton.org, R2
- le site http://recettessimples.fr, R3

Deux types d'informations ont été dégagés : les métadonnées (Tableau 8) et les opérations mentionnées dans la préparation d'une omelette (Tableau 9).

	Françoise Bernard	**marmiton.fr**	**recettessimples.fr**
Temps de préparation	5 min	5 min	
Temps de cuisson	6 min	10 min	
Ingrédients	9 œufs 50 g de beurre ¼ de verre d'eau Sel, Poivre	7 œufs 50 g de beurre Sel, poivre	4 œufs Beurre Sel, poivre
Nombre de personnes	6	4	

Tableau 8 : Métadonnées

Cette comparaison a permis de reconstituer la liste des ingrédients et les principales opérations en distinguant d'un côté les étapes obligées (*casser les œufs etc.*) et les variations admises dans la réalisation (*degré de cuisson, assaisonnement etc.*). Six opérations de base ont été dégagées : *casser les œufs, battre, saler/poivrer, beurre/chauffer la poêle, verser les œufs dans la poêle, cuire.* Nous avons observé pourtant un certain nombre de variations entre les énoncés (Tableau 10). Cette disparité dans l'uniformité se retrouve aussi dans l'ordre des actions qui admet certaines variantes significatives, autant par les ellipses (si on bat les œufs, c'est qu'ils sont cassés ; si on les verse dans la poêle, c'est pour les faire cuire) que par la rigidité du déroulement. Les six actions principales ont été énumérées pour pouvoir comparer et analyser leur ordre (Tableau 11) :

1 casser les œufs
2 les battre

[38] Il s'agit du livre *Recette faciles* édité en 2008 par Hachette cuisine.

3 saler (poivrer)
4 chauffer dans la poêle la matière grasse
5 *verser les œufs battus*
6 *(faire) cuire*

Etapes	Françoise Bernard	marmiton.org	recettessimples.fr
1.	Cassez les œufs dans un petit saladier	Battez les œufs à la fourchette	Dans un bol casser les œufs
2.	Salez, poivrer	Salez et poivrez	Ajouter du sel, du poivre
3.	Versez de l'eau	Faites chauffer le beurre	Battre et mélanger avec une fourchette
4.	battez vivement le mélange sans insister trop longtemps	Versez-en un peu dans les œufs	Dans une poêle adaptée, faire fondre un morceau de beurre
5.	Mettez le beurre dans une grande poêle	Mélangez	Lorsque le beurre est fondu, ajouter les œufs battus
6.	Lorsqu'il commence à roussir Versez-y les œufs battus	Versez les œufs dans la poêle à feu vif	Remuer au début avec une cuiller en bois
7.	Détachez les bords de l'omelette en passant une fourchette entre les œufs et la poêle, tout autour. De temps en temps ramener le bord de l'omelette vers le centre avec la fourchette.	Baissez le feu	Lorsque l'omelette est presque cuite, mais encore "baveuse", replier une moitié sur l'autre dans la poêle
8.	Laissez cuire 5 à 6 minutes Soulevez le bord : il doit être doré tandis que le dessus de l'omelette est encore baveux.	Laissez cuire doucement en ramenant les bords de l'omelette au centre au fur et à mesure qu'ils prennent	Servir
9.	Inclinez la poêle au-dessus d'un plat et à l'aide d'une fourchette roulez l'omelette et faites-la glisser sur le plat. Servez très chaud	Pliez l'omelette en deux	
10.		Servez	

Tableau 9 : Opérations

Il est intéressant de noter que contrairement aux variations lexicales observées, on constate très peu de modifications dans l'ordre du déroulement de la recette : la recette proposée par le

site R2 fait commencer la recette à la battue, les deux autres salent et poivrent au début de l'opération. Ce sont les seules permutations observées.

Actions	**Françoise Bernard**	**marmiton.fr**	**aufeminin.com**
Casser	***cassez les œufs dans*** un petit saladier		***dans*** un bol, ***casser les œufs***
Battre	**battez** vivement le ***mélange*** sans insister trop longtemps	**battez** les œufs à la ***fourchette***	**battre** et ***mélanger*** avec une ***fourchette***
Saler, poivrer	**salez**, **poivrer**	**salez** et **poivrez**	ajouter du **sel**, du **poivre**
Beurre, chauffer la poêle	mettez **le beurre *dans une*** grande **poêle**	faites chauffer **le beurre**	***dans une*** **poêle** adaptée, faire fondre un morceau de **beurre**
Verser les œufs dans la poêle	lorsqu'il commence à roussir **versez**-y **les œufs *battus***	**versez les œufs dans** la **poêle** à feu vif	lorsque le beurre est fondu, ajouter **les œufs *battus***
Cuire	détachez ***les bords de l'omelette*** en passant une fourchette entre les œufs et la poêle, tout autour. De temps en temps ramener le bord de l'omelette vers le centre avec la fourchette.	laissez ***cuire*** doucement en ramenant ***les bords de l'omelette*** au centre au fur et à mesure qu'ils prennent ***pliez*** l'omelette en deux	remuer au début avec une cuiller en bois. Lorsque l'omelette est presque ***cuite***, mais encore "baveuse", ***replier*** une moitié sur l'autre dans la poêle. Servir

Tableau 10 : Variation lexicale des actions principales dans les trois recettes[39]

Françoise Bernard	**marmiton.org**	**recettessimples.fr**
1	2	1
3	3	3
2	4	2
4	5	4
5	6	5
6		6

Tableau 11 : Ordre des opérations dans les trois recettes

[39] Les éléments répétés ou synonymes sont mis en gras s'ils sont présents dans les 3 recettes, en gras et italiques s'ils sont seulement communs à deux d'entre elles.

Partant de cette observation des trois exemples-modèle des recettes écrites, on peut distinguer approximativement ce qui relèverait d'une recette type. Nous émettons l'hypothèse que les recettes proposées dans le cadre de l'interview doivent contenir en plus des informations classiques propres aux recette, d'autres types d'informations appartenant à la communication orale comme les marques de l'énonciation, les éléments phatiques, la diversité des niveaux du discours (interpellations de l'auditeur, anecdotes etc.), les connecteurs ou encore les disfluences.

3.6.4.1. Annotation

Pour pouvoir procéder à l'analyse des informations présentes, celles-ci devaient être annotées. Une observation manuelle du corpus a permis de distinguer quatre types d'informations :

- *les actions élémentaires* correspondant aux six étapes rudimentaires de la recette typique de l'omelette ;
- *les expansions* concernant les éléments secondaires liés à la recette ;
- *les développements latéraux* regroupant des données qui entretiennent un rapport indirect avec la recette ;
- *les méta-commentaires* concernant la situation de communication.

Ces informations ont été annotées comme dans l'exemple :

<entretien nr="006">

cette euh cette dernière question euh je crois que la question va vous faire rire mais je la pose quand même et

* <mc prec="formule de politesse">hm je vous en prie</mc> *

* <mc prec="commentaire">il y a pas de danger oh la</mc> *

hm

je voudrais vous demander comment est-ce qu'on fait une omelette chez vous

* <mc prec="réaction première">une omelette</mc> *

une omelette

* <action> en cassant les oeufs</action> *

Le format XML de l'annotation a permis l'extraction de ces informations grâce aux feuilles de style XSLT sous forme de tableau (Tableau 12). Les résultats de cette extraction ont été analysés.

3.6.4.2. Méta-commentaires

Les métadonnées dans les recettes concernent : les réactions des interviewés à la question, les anecdotes de la vie courante liées indirectement à la recette, les capacités de cuisiner que les locuteurs affirment avoir ou non et la conclusion de la recette.

3.6.4.2.1. Réactions à la question posée

La question relève de la série des questions posées au cours de l'entretien qui portent sur le locuteur lui-même, ses origines, son travail, sa famille, ses loisirs ou encore sur le français parlé à Orléans. Les différentes stratégies mises en œuvre par les intervieweurs pour introduire « en douceur » cette question ont été analysées dans (Abouda et Perrot 2006).

Je me contenterai ici de présenter les différentes réactions de la part des interviewés à cette question, liées à l'originalité de la question et à la surprise qu'elle suscite. Une compétence culinaire ne peut s'apprécier à partir d'une préparation aussi simple. Tous les enquêtés ont sur le sujet une connaissance pratique qui est du même ordre que celle des enquêteurs. La réaction des interviewés est donc variable.

Numéros d'entretiens	Actions élémentaires	Expansions	Développements latéraux	Méta-commentaires
008	# on casse les œufs # on bat tout ensemble # on assaisonne sel # on verse tout ça dans la dans la poêle et puis on tourne jusqu'à temps que ça soit à peu près cuit quoi	# on met un peu d'eau j(e) crois # on mélange un peu d'eau enfin # poivre # moi je on met des des épices aussi dedans quand même quelquefois des de l'ail de l'oignon	**recette** oh on peut la faire de plusieurs façons **recette** oui il y a plusieurs façons i(l) y a plusieurs façons hein i(l) y a plusieurs façons **recette** oh on la fait de plusieurs façons **régions-pays** en LORRAINE on on découpe des petits des petits morceaux de lard qu'on fait frire avant **régions-pays** seulement ici on ne l'a fait pas au lard parce qu'on ne trouve pas de de charcuterie comme en comme en LORRAINE **régions-pays** on trouve non euh la charcuterie ici c'est pas très fort et bah on n'a qu'a du lard fumé des des saucisses fumées vous savez c'est pas du tout fait pareil ici enfin bon	**commentaire** elle allait en faire une justement c'est ce qu'on va manger ce soir **commentaire** mais enfin je sais quand même faire une omelette **commentaire** je m'en suis fait moi-même aussi hein

Tableau 12 : Extrait de l'extraction des données annotées

- répétitions :

  *DP : **comment** est-ce qu'**on fait une omelette** chez vous*

  *FC 716 : **comment on fait une omelette** et bien **comment on fait une omelette*** (66B)

  *JR : euh voulez-vous me raconter ce que vous feriez **si vous étiez seul à la maison si vous aviez très très faim et vous deviez préparer une omelette***

  *GS : **si j'étais seul à la maison si j'avais très très faim et si je me préparais une omelette*** (75B)

- hésitation : le locuteur ne comprend pas la question et en demande confirmation. On retrouve aussi les disfluences :

  ***alors euh** ensuite il faut que je vous dise tout après comment il faut faire* (63B)

ah *c'est la recette que vous voulez* ***et la*** *et la technique* ***bon euh et bien euh*** (144B)

- justification de la question : le locuteur souligne le fait que la question coïncide avec son expérience personnelle, que la question lui convient.

  ah vous ***tombez bien c'est moi qui les fait*** *toutes ici* (6B)

  elle allait en faire une justement *c'est ce qu'on va manger* ***ce soir*** (8B)

  j'ai appris ça à mon fils *cadet il y a très il y a peu de temps* (82B)

- différentes variétés d'omelettes : le locuteur précise d'emblée qu'il existe plusieurs façons de faire une omelette et qu'il doit donc choisir. Le nom *façon* et ses synonymes, des mots exprimant la quantité ou encore l'expression *ça dépend* reviennent dans les réponses :

  vous avez des ***différentes façons*** *de la faire* (9B)

  oui bah euh je parce que non mais parce qu'il y a ***trente-six façons*** *je veux dire on peut faire des omelettes aux champignons au jambon au non une omelette simplement une omelette* (60B)

- simplicité de la recette : une autre introduction de la recette passe par sa qualification :

  eh bien ***c'est très simple*** (64B), ***c'est pas très difficile*** (106B)

Cette simplicité peut être mentionnée aussi dans la recette à travers le temps de préparation :

c'est ***très vite fait*** (131B)

oh ***ce n'est pas long*** *une omelette une omelette* ***quatre cinq minutes*** (124B)

c'est ***l'espace de deux trois minutes au plus*** *c'est tout* (22B)

- irritation : la question peut susciter l'étonnement, voire l'irritation du locuteur :

  ah vous m'en posez des questions *vous ah vous m'avez eu là* (83B)

  une omelette et ***pourquoi vous me demandez ça*** (94B)

  eh bien ***mon Dieu*** (109B)

3.6.4.2.2. Capacités de cuisiner

Les locuteurs embarrassés parfois par la question posée mettent en doute leurs capacités de cuisiner :

ça m'arrive rarement d'en faire parce que ***je ne suis pas très très attiré sur le plan cuisine*** (64B)

c'est une question plutôt embarrassante pour moi parce que vous savez ***au point de vue cuisine*** *vous savez* ***je n'en fais pas beaucoup*** (111B)

Certains refusent ou détournent la réponse

ah ***je n'en sais rien*** *je n'en sais je n'en sais rien* (97B)

c'est assez difficile de vous expliquer *parce que la cuisine euh à part faire cuire un beefsteak et des frites c'est tout ce que je sais faire* (130B)

Ces réponses sont liées en réalité non à une ignorance mais à une inhibition ou à des représentations de soi tacites. Ce sont les cas où les hommes craignant de déroger à leur statut viril renvoient la question vers leur épouse :

oh bien écoutez alors là vous savez moi pour faire une omelette euh ***faudra plutôt vous intéresser à ma femme*** (111B)

vous savez que je m'intéresse ***j'ai une femme cuisinière*** *et que je me suis bien gardé de d'y mettre mon nez je lui laisse le soin de le faire* (122B)

On retrouve aussi les cas inverses où ce sont les femmes qui délèguent la parole à leur mari :

mais ***mon mari*** *vous savez* ***est très bon cuisinier*** *il peut vous répondre* (133B)

je vais vous dire ***ce n'est pas moi qui la fait c'est mon mari*** (22B)

Il est possible que ce soit la nature du plat considéré comme très simple qui ait déterminé ces réponses.

D'autres locuteurs, en éludant la banalité de la recette, veulent se démarquer et disent qu'ils la font à leur manière en insistant de cette façon sur l'originalité de leur pratique :

je fais l'omelette un peu ***à ma manière à moi*** (14B)

eh bien on n'est ***pas toujours obligé de la faire de la même façon*** (21B)

Certains locuteurs veulent justifier leurs talents de cuisine en donnant des précisions sur les événements, occasionnels ou fréquents, où ils ont pu faire la preuve de leur talent :

je faisais ça ***parce que*** *j'étais* ***longtemps*** *c'est* ***moi qui me faisais ma cuisine*** *tout seul* (14B)

parce que *je suis quand même un spécialiste des questions de cuisine […]* ***l'un de mes passe-temps favoris*** *c'est de faire la cuisine (78B)*

quand j'invite des amis *à déjeuner les questions de cuisine me sont intégralement réservées* (78B)

ah ben ***je suis forcée de faire ma cuisine des fois*** *je rentre huit jours de vacances plus tôt que ma femme […] alors pendant huit jours ben je fais mon petit frichti moi-même* (47B)

En effectuant une recherche sur le mot *cuisine* employé 29 fois, on constate qu'il est toujours utilisé pour justifier ses aptitudes en cuisine, qu'elles soient ou non avérées.

3.6.4.2.3. Anecdotes et plaisanteries

On peut observer, au nombre des méta-commentaires des interviewés, des anecdotes ou des remarques destinées à faire rire l'interlocuteur :

d'ailleurs on dit en France qu'une femme c'est comme une omelette plus elle est battue meilleure elle est (21B)

mh oui oui vous aimez la cuisine française [rire] nous sommes des gourmands nous [rire] (106B)

alors je ne répondrais pas sans casser les œufs (133B)

3.6.4.2.4. Fin de la recette

Les locuteurs utilisent différentes stratégies pour marquer la fin de la recette :

- puisqu'il s'agit d'une énumération d'étapes, les personnes terminent souvent leur recette par des mots et expressions de conclusions :

voilà (43 occurrences), *c'est tout* (27 occurrences), *ça y est* (4 occurrences), *et zou* (1 occurrence) etc.

- certains commentent leur recette par l'action finale, *manger :*

  vous allez pouvoir manger des omelettes à la façon française (136B)

- d'autres ajoutent un commentaire sur le goût : *c'est bon, c'est délicieux etc.*

3.6.4.3. Développements latéraux

Les développements latéraux concernent les données qui n'entretiennent pas un rapport direct avec la recette. Il s'agit de commentaires personnels ou de conseils liés à la recette :

> *seulement ici on ne l'a fait pas au lard parce qu'on ne trouve pas de de charcuterie comme en comme en Lorraine […] on trouve non euh la charcuterie ici c'est pas très fort et bah on n'a qu'à du lard fumé des des saucisses fumées vous savez c'est pas du tout fait pareil ici enfin bon* (008B)

> *vous avez l'omelette a- avec euh des oignons et vous avez l'omelette comme ça nature et vous avez l'omelette avec euh la crème fraîche puis vous avez encore les omelettes euh aux asperges les omelettes avec différents légumes* (009B)

3.6.4.4. Expansions

Les expansions sont les éléments supplémentaires enrichissant la recette type. Il peut s'agir

- des ingrédients complémentaires :

  je mets des ***tranches de bacon*** *dedans* (006B)

  en Lorraine on on découpe des petits des ***petits morceaux de lard*** *qu'on fait frire avant* (008B)

- des ustensiles :

  on prend une ***poêle*** *Tefal* (22B)

  en battant des œufs avec une ***fourchette*** (46B)

Chaque action principale correspond à un ou plusieurs ustensiles appropriés : *casser les œufs* dans un *saladier, bol* etc. ; *les battre* avec une *fourchette, un fouet* etc. ; *mettre une matière grasse* et *verser* tout dans un *poêle ;* etc. Ainsi, pour chaque étape, une catégorie d'objets spécifiques est mentionnée.

- des précisions quantitatives

  suivant le goût peut-être ***un petit peu de poivre*** *pour ceux qui l'aiment et* ***une cuillerée à bouche*** *d'eau ça rend plus léger* (26B)

  je casse euh ***deux trois œufs*** *comme ça et puis je les bats pendant* ***trois quatre minutes*** *quoi* (60B)

- des précisions sur la manière d'effectuer une opération de base :

  je ***bats très fortement*** *mes œufs très très fortement de façon à ce qu'ils moussent bien* (21B)

  vous ***salez*** *et vous* ***poivrez légèrement*** (144B)

  je la laisse ***cuire doucement*** (135B)

Le plus souvent les locuteurs insistent sur l'intensité d'une opération *battre les œufs* ou *cuire*. On trouve également des variations lexicales :

battre/mélanger/remuer/fouetter/agiter/débattre/brouiller/secouer

fortement/bien/énormément/vivement/longtemps/doucement

pas trop/finement/légèrement/le moins possible

- des précisions temporelles

sur l'ordre des opérations :

puis quand *je vois qu'elle est bien faite et bien je la mets dedans* ***puis*** *je la fait retourner de temps en temps avec une fourchette je je gratte un petit peu* ***c'est tout*** (16B)

d'abord *on bat les œufs bon* ***alors*** *ça cette chose étant bien battue je mets du beurre dans le dans la poêle* (21B)

sur le temps de préparation

peut-être ***sept minutes*** (15B), *c'est l'espace de* ***deux trois minutes au plus*** *c'est tout pour cuire une omelette* (22B) etc.

- des justifications de certaines opérations

actuellement ***avec notre régime*** *on ne met pas de beurre on prend une poêle TEFAL* (22B)

j'ajoute un peu de lait ***parce que*** *ça fait une omelette plus légère* (94B)

mais je je pique de temps en temps comme ça pour euh ***pour*** *pas qu'elle ne colle* (96B)

Les locuteurs donnent aussi leur avis concernant la recette. Ces expansions sont introduites souvent par des verbes cognitifs *penser, croire, trouver* etc., des verbes de sentiments *aimer etc.* ou encore des expressions *à mon avis, pour moi* évoquant une opinion personnelle.

*on ne l'****aime*** *pas trop cuite* (102B)

parce que ***je trouve que*** *quand on les bat de trop ils se tournent en neige ils ne sont il n'en est pas meilleur* (147B)

enfin en principe une omelette ***à mon avis*** *pour qu'elle soit bon il faut qu'elle soit ce qu'on appelle baveuse* (45B)

- des doutes sur certaines opérations.

Le vocabulaire mentionné dans le point précédent peut être utilisé aussi pour indiquer les doutes, les hésitations :

euh ***je crois que*** *d'abord euh mettre un petit de beurre* (120B)

je ne sais pas *euh pour euh quatre personnes euh je mettrais un œuf ou deux en plus* (136B)

je suppose *il faut mettre du beurre déjà* (130B)

- des précisions sur la température :

servez ***très chaud*** (78B)

on fait chauffer à ***feu vif*** (156B)

- des précisions sur la vérification de l'état des ingrédients :

  *quand le beurre est **fondu*** (156B)

  AR:comment est l'omelette […]est-ce qu'elle change est-ce qu'elle reste identique est-ce qu'elle change de couleur

  […]

  RV 252:elle roussit dessous (123B)

- des actions supplémentaires pendant la cuisson

  *je laisse cuire mais je **je pique de temps en temps*** (96B)

  *euh en **en ramenant euh régulièrement** euh le euh la part de d'œuf qui est cuit sur le fond de le de la poêle au centre de la poêle **pour feuilleter*** (156B)

- des précisions sur la façon de servir le plat

  ***vous la retournez en sabot** dans votre plat pour servir* (009B)

  ***vous la repliez en deux** vous la servez en principe pas trop cuite* (26B)

  *vous versez l'omelette euh jusque la moitié dans une assiette et vous retournez très rapidement la poêle de façon à **la servir pliée en deux comme un chausson*** (144B)

3.6.4.5. Actions élémentaires

L'analyse des trois recettes écrites prises en référence a permis de distinguer six opérations fondamentales nécessaires à la préparation d'omelette : *casser les œufs, les battre, saler (poivrer), chauffer dans la poêle la matière grasse, verser les œufs battus, (faire) cuire*.

Ces opérations, leurs variations et leur ordre d'enchaînement ont été examinées.

3.6.4.5.1. Analyse lexicale

Chaque opération fondamentale est enrichie par des expansions qui lui sont propres :

- *casser les œufs* : le nombre d'œufs qu'on utilise dans la recette.
- *battre les œufs* : l'intensité forte de l'opération et sa durée (*longtemps, fortement, de façon à ce qu'ils moussent bien etc.*).
- *saler (poivrer)* : l'intensité faible (*un petit peu, légèrement etc.*).
- *chauffer la matière grasse* : la quantité de matière grasse et l'ustensile (*Tefal, poêle*).
- *verser les œufs* : cette action suivant immédiatement la précédente exige l'application d'une certaine condition (*la poêle doit être chaude, la matière grasse doit être fondue etc.*).
- *faire cuire* : le type de cuisson (*pas brunir dorer*), la façon de le faire (*sans trop remuer*). On note plusieurs fois l'emploi du verbe *dorer* qui qualifie le type de cuisson et le terme *baveuse* qui qualifie la cuisson finale.

3.6.4.5.2. Analyse quantitative

Le Tableau 13 présente quelques chiffres de fréquence des opérations fondamentales dans le corpus. On voit que chaque opération est mentionnée dans plus de 50 % des cas. La plus répandue semble être *battre les œufs* alors que la moins fréquente est *cuire*. Les locuteurs omettent parfois cette dernière opération car ils concluent la recette par *verser les œufs battus dans la poêle* ou la remplacent par des précisions sur le service.

3.6.4.5.3. L'ordre des opérations fondamentales

En se basant sur les recettes de référence, chaque opération fondamentale a été codée par un chiffre indiquant l'ordre de son apparition dans la recette :

1 *casser les œufs*
2 *les battre*
3 *saler (poivrer)*
4 *chauffer dans la poêle la matière grasse*
5 *verser les œufs battus*
6 *(faire) cuire*

	casser les œufs	**battre**	**saler, poivrer**	**chauffer la poêle**	**verser dans la poêle**	**cuire**
occurrences	77	92	59	68	74	50
pourcentages	81,9%	97,8%	62,7%	72,3%	78,7%	53,1%

Tableau 13 : Occurrence des actions

Le Tableau 14 et la Figure 4 montrent la fréquence de l'ordre d'apparition de chacune des opérations dans le corpus. On notera que l'ordre défini par les recettes de référence n'est pas toujours repris à l'oral. Ainsi, le locuteur peut commencer par n'importe quelle opération avant de remonter vers la première et de retourner à l'ordre attendu.

	1	2	3	4	5	6
casser	67	8	2	0	0	0
battre	13	42	29	7	1	0
saler	1	27	18	7	5	1
chauffer la poêle	13	8	20	20	6	1
verser dans la poêle	0	5	11	34	23	1
cuire	0	0	4	8	17	21
totale	94	90	84	76	52	24

Tableau 14 : L'ordre de l'apparition de chaque opération.

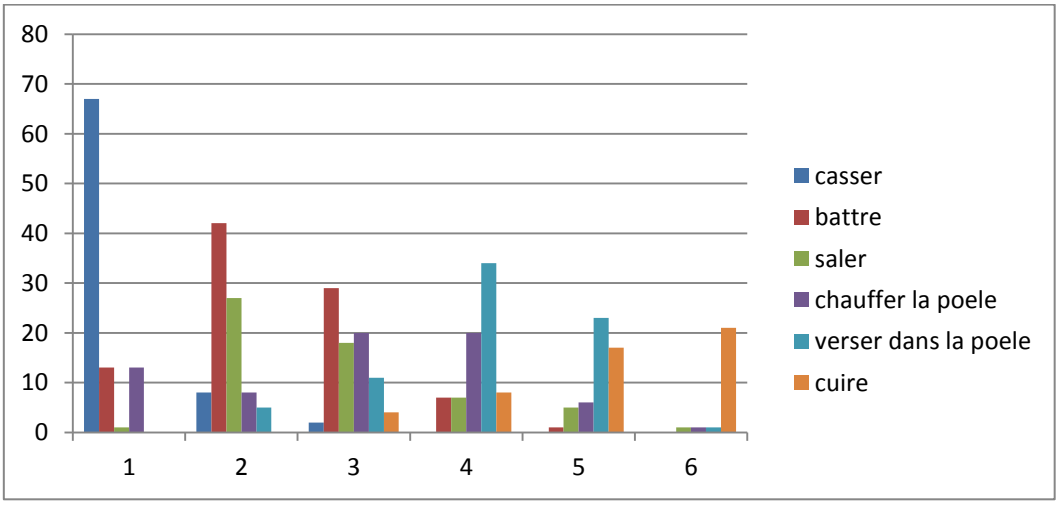

Figure 4 : Distribution de chaque opération selon son ordre d'apparition

Il peut omettre une action en la rendant implicite (comme c'est le cas de *casser les œufs, verser dans la* poêle) ou en l'oubliant.

Les opérations peuvent se suivre dans un ordre assez aléatoire, s'inverser, être omises ou être répétées comme dans l'exemple :

> *JSM: est-ce que vous pourriez me décrire m'expliquer comment on fait une omelette ?*
>
> *1300: dans une poêle on du beurre on met du beurre avant oui mais avant on a battu des œufs on a cassé des œufs on a battu des œufs avec du sel du poivre*
>
> *JSM: mh mh*
>
> *1300: euh quand le beurre est fondu on met le tout dans une dans la poêle on fait cuire on fait chauffer à feu vif euh en en ramenant euh régulièrement euh le euh la part de d'œuf qui est cuit sur le fond de le de la poêle au centre de la poêle pour feuilleter et puis on laisse prendre un petit peu au dernier instant et puis on démoule enfin on verse* (156B)

L'Annexe 14 montre l'ordre d'apparition des opérations trié en fonction du code et de la fréquence. Pour les codes, 13 signifie que le locuteur est passé directement de la première à la troisième action, 01 qu'il n'y a rien avant la première et 60 que la recette se termine sur la sixième. Si *casser les œufs* et *les battre* s'imposeraient naturellement dans cet ordre (codé 12), 13 témoins s'autorisent à commencer l'omelette à partir du moment où ils *battent les œufs*, partant de cette évidence que l'opération ne peut être réalisée qu'une fois les œufs cassés. Une dizaine de locuteurs revient par après sur cet oubli et réintroduit ultérieurement l'opération comme le montrent, dans le codage réalisé, un schéma du type 41 qui signifie que le corps gras a été déposé dans la poêle avant qu'il ne soit fait mention que les œufs ont été cassés.

Une autre façon de visualiser les résultats se trouve dans les Figure 5 et Figure 6. Dans la Figure 5 sont regroupées trois types de fréquences sous forme de flèches ce qui permet de mettre en évidence certaines tendances observées dans l'ordre d'enchaînement entre les opérations :

- les ordres les plus fréquents sont : *12, 23, 32, 13, 24, 25, 45, 56, 50*. Il s'ensuit que *caser les œufs* est suivi par *battre* ou *saler (poivrer).* L'action de *battre* qui remplace dans certains cas celle de *casser* peut être suivie de trois autres : *saler ou poivrer, chauffer la matière grasse, verser dans la poêle.* Les deux opérations se permutent souvent *battre* et *saler.* La recette se finit souvent par *verser les œufs dans la poêle* et/ou par *cuire* ;
- les ordres rares, dont la fréquence est limitée à 5 occurrences ou mois : *21, 26, 62, 03, 36, 63, 43, 14, 54, 15.* Il est rare que le locuteur *casse les œufs* avant de *les battre,* qu'il *cuise les œufs* avant *les avoir battus* ou *salés.* Pourtant à l'oral cet ordre « bizarre » peut apparaître dans les cas où le locuteur se rappelle avoir oublié de mentionner une action. Parfois, le locuteur commence la recette directement par *saler (poivrer).*
- les ordres non attestés dans le corpus : *16, 61, 31, 51, 05, 06, 52, 65.* Il est évident qu'on ne peut pas commencer la recette par *cuire* ou *verser dans la poêle* qui sont deux actions anaphoriques. Ainsi, *battre* ne peut pas suivre l'action de *verser dans le poêle* et *saler (poivrer)* ne peut permuter avec *casser.* De la même manière, l'opération de *casser les œufs* ne peut apparaître après *saler, verser* et *cuire.*

La Figure 6 montre la même visualisation mais en prenant en compte le sexe du locuteur. Les recettes féminines semblent respecter plus volontiers l'ordre classique entre opérations.

3.6.4.6. Différence entre les recettes de référence et les recettes du corpus

Si l'on compare les recettes de référence et celles de notre corpus, on distingue deux types d'informations distinctes :

- des informations communes à toutes les recettes et qui sont inhérentes à ce genre de discours : actions élémentaires et expansions concernant les ustensiles, temps de cuisson, quantité d'ingrédients etc.
- des informations propres à l'oral liées à la situation de communication, au dialogue, aux critères sociologiques des locuteurs etc.

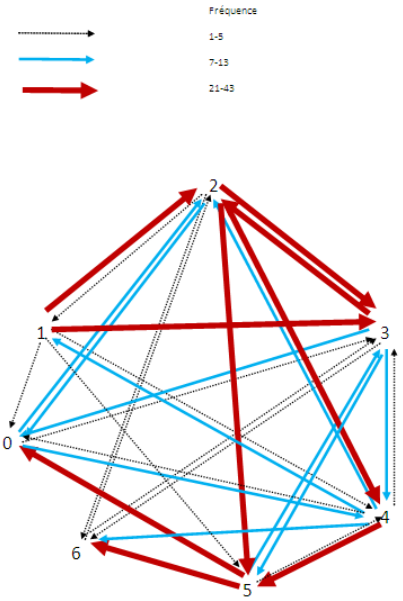

Figure 5 : Distribution de l'ordonnancement entre les actions élémentaires sous forme de graphe

Soit l'entretien 66B

> *DP: alors finalement euh pourrez-vous me dire euh comment est-ce qu'on fait une omelette chez vous ?*
>
> *FC 716: comment on fait une omelette ? eh bien comment on fait une omelette euh très facile euh* ***on prend le la poêle*** *hein* ***on met du beurre*** *euh* ***certains mettent un petit peu d'huile on fait bien chauffer*** *ça* ***on casse les œufs on les bat***
>
> *DP: mh ah*
>
> *FC 716: puis ben ma foi* ***on met ça dans la poêle*** *et* ***je remue la poêle*** *avec euh pour la rouler* ***avec*** *euh* ***une fourchette***
>
> *DP: oui*
>
> *FC 716:* ***je cogne légèrement sur la la queue de la poêle pour la décoller*** *et* ***l'omelette se trouve roulée*** *et* ***on verse dans un plat***

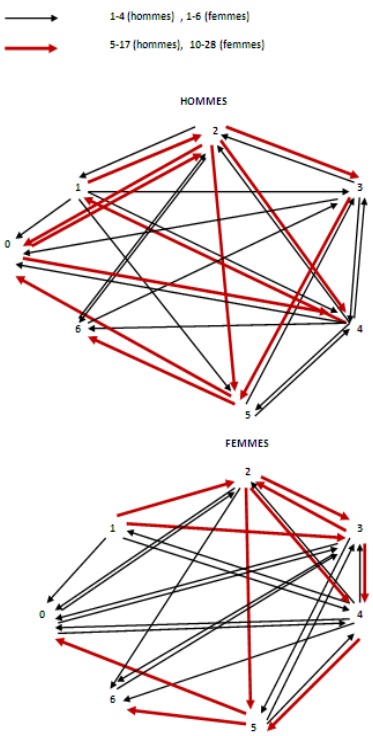

Figure 6 : Distribution de l'ordonnancement entre les actions élémentaires sous forme de graphes en fonction du sexe du locuteur

On y retrouve les opérations élémentaires de confection de l'omelette mais leur ordre paraît aléatoire. A cela s'ajoutent des commentaires sur la façon d'effectuer certaines de ces opérations ou des actions supplémentaires. La recette est enrichie également par les éléments propres à l'oral comme les hésitations ou encore les interjections qui montrent l'aspect personnel de la réponse. Le locuteur n'assume pas au début sa réponse et emploie le *on* avant de s'affirmer en première personne. En dialogue, la présence de l'interlocuteur influence la conduite. Le locuteur s'est étonné de la question, prenant un certain temps avant de formuler sa réponse. Les mêmes procédures peuvent être observées dans l'entretien 132B :

> *tout d'abord on casse les œufs dans un saladier on mélange le blanc et le jaune et puis on bat avec soit avec une fourchette soit avec un appareil parce qu'il existe des appareils pour battre les œufs et euh vous prenez une poêle vous y mettez du beurre et quand il est fondu euh vous versez l'omelette par-dessus j''ai oublié de dire qu'on mettait du sel et du poivre dans les œufs*

où le locuteur introduit un commentaire au cours de la recette sur des mixeurs industriels, permutant les opérations élémentaires et alternant *on*, *vous* et *je*.

On note que les commentaires sont très variés et qu'ils s'éloignent parfois de la question de départ ; aussi que les gens ressentent le besoin de préciser nombre d'éléments de la recette ou de leur vie privée pour justifier leurs propos. Parfois, ils semblent pris de court, hésitent à répondre, comme s'il leur fallait préparer leur discours. Leur réponse concerne moins la préparation en tant que telle que le rapport que les personnes entretiennent à la cuisine.

3.6.5. Bilan

3.6.5.1. Compte rendu

Objectif visé	Comparer deux types de recettes : les recettes orales proposées dans le cadre d'un dialogue et les recettes classiques des manuels ou des sites culinaires ; Analyser quantitativement et qualitativement les informations contenues dans le corpus oral des recettes de préparation d'omelettes.
Méthodologie	Utilisation des outils de linguistique de corpus : concordancier, feuille de style XSLT, tableur Excel pour pouvoir annoter, extraire et analyser les informations annotées.
Données traitées	96 réponses à la question posée au cours des entretiens en face-à-face : Comment est-ce qu'on fait une omelette ? Pourriez-vous m'expliquer comment on fait ? Pouvez-vous me donner la recette de l'omelette ?
Difficultés rencontrées	Des caractéristiques propres aux corpus oraux transcrits
Résultats	Corpus : 96 réponses annotées entiérement Analyse quantitative et qualitative des variations de l'information annotée Elaboration de la typologie des informations : *actions élémentaires, expansions, développements latéraux, méta-commentaires* Etude des variations et de l'ordre d'enchaînement
Contraintes et limites	Non prise en compte des informations prosodiques
Originalité et apport du travail	Les recettes de cuisine écrites ont fait l'objet de différentes études, mais il n'en va pas de même avec des recettes transmises oralement. Comparaison entre deux types de recettes : les recettes orales proposées dans le cadre d'un dialogue et les recettes classiques des manuels ou des sites culinaires. Dans le premier cas, on assiste à l'élaboration de la recette, d'où les permutations ou les oublis des actions élémetaires, les commentaires liées ou non à la recette, l'expression de sentiments et d'opinions, les refus de réponse, l'interaction avec l'enquêteur etc. Dans le deuxième cas, l'auteur de la recette suit un format prédéfini de la recette. La recette de cuisine livrée à l'oral dans le cadre des entretiens a une connotation subjective qui ne se retrouve pas dans les recettes proposées dans les manuels ou sur le Web. Phénomènes et variations linguistiques très intéressantes issues de l'étude du corpus de recettes d'un plat aussi simple qu'une omelette : - son mode oral de transmission et de restitution ; - une réponse à une question posée dans un cadre inattendu, celui d'un dialogue où la présence de l'interlocuteur a été prise en compte dans l'analyse.

3.6.5.2. Perspectives et travaux en cours

La suite des travaux peut concerner l'étude diachronique en comparant les réponses d'ESLO1 à celles d'ESLO2. On ne s'attend pas à une différence dans la recette elle-même en supposant qu'elle n'a pas évolué dans le temps. Les actions élémentaires doivent donc rester identiques. La différence attendue concerne plutôt *les expansions, les développements latéraux et les méta-commentaires*.

L'étude proposée va au-delà de la linguistique du corpus. Ses résultats peuvent être appliqués au domaine de la représentation des connaissances pour générer des recettes personnalisées avec une extension aux applications du dialogue homme-machine, de la synthèse vocale etc.

Le travail présenté dans cette partie est le dernier dans la série des travaux concernant l'annotation du corpus ESLO. Tous ces travaux suivent le principe du respect des données traitées. Je travaille avec les méthodes symboliques comme avec les méthodes d'apprentissage automatique. Le contenu de l'annotation et les méthodes automatiques utilisées varient en fonction des tâches à effectuer mais la démarche reste toujours la même : je pars du corpus afin d'observer les phénomènes que j'entends annoter, ce qui me permet de définir le jeu d'étiquettes et la technique pour l'automatisation du processus. Le corpus annoté permet de faire des analyses quantitatives et qualitatives. Ainsi, l'oral rend subjectives certaines unités ou certaines informations qui sont par nature plutôt neutres.

4. Noms de lieux sur le Web

Ce nouveau chapitre est consacré à des recherches sur le corpus écrit composé des titres de cartes issu du Web. Elles consistent dans le repérage, l'annotation et l'analyse de l'information spatiale dans ce corpus. Une fois de plus, il s'agit d'un corpus non standard, une sorte de combinaison entre discours oral et écrit, qui exige un traitement spécifique. Ce travail a été effectué en collaboration avec Catherine Domingues, chargée de recherches à l'Institut de l'information géographique et forestière (IGN).

4.1. Résumé du travail

La notion de *lieu* et son identification posent différents problèmes. Le nommage des objets géographiques n'est pas normalisé et vient souvent de la tradition orale. La nature des noms de lieu varie : noms propres, noms communs, déictiques. Un endroit peut faire référence à un lieu imaginaire ou métaphorique. Les noms de lieux peuvent être ambigus : un même lieu peut être désigné par plusieurs toponymes et inversement un même toponyme peut désigner des objets géographiques différents. Du point de vue typographique, ils peuvent commencer ou non par une majuscule. Comme la désignation d'un lieu est un processus social réapproprié subjectivement, la signification d'un toponyme, c'est-à-dire la référence à un lieu donné, est déterminée par le système toponymique du locuteur.

L'objectif du travail a été de repérer et analyser les désignations actuelles de lieu par des locuteurs variés sur le Web. Le corpus traité riche en désignations de lieux a été tiré du Web. Il s'agit d'un corpus des titres de cartes géographiques.

Une analyse détaillée du corpus a été effectuée et la typologie des lieux a été définie. Pour repérer les désignations de lieux dans le corpus, nous avons utilisé la méthode symbolique utilisant les grammaires locales. La méthode proposée s'appuie ainsi, d'une part, sur l'utilisation des ressources lexicales contenant les toponymes français et d'autre part, sur la description du contexte de l'emploi du nom de lieu sous forme de patrons. Nous avons ainsi annoté 50942 titres.

Plusieurs des spécificités du corpus traité ont rendu son traitement automatique difficile :

- les conventions de nommage et d'écriture y sont fluctuantes et varient d'un utilisateur à l'autre ;

- les titres des cartes ont un caractère personnel et parfois intime ;

- tous les titres ne sont pas en français.

- le titre peut être complexe et contenir des sous-titres mais sans usage systématique de marques de segmentation.

- les titres sont livrés sans contexte.

L'analyse des variations dans la désignation de lieux et dans l'écriture de ces lieux sur le Web par les différents utilisateurs ainsi que l'étude de la perception des lieux et des sentiments chez l'utilisateur ont été effectuées. Deux phénomènes en particulier ont été mis en évidence : le non respect des règles d'écriture des noms de lieu et la présence des lieux appelés *subjectifs*.

Le travail sur le corpus des titres de cartes géographique a été décrit dans (Domingues et Eshkol-Taravella 2013, à paraître a,b).

4.2. Qu'entend-on par lieu ?

La notion de lieu s'appuie sur des définitions hétérogènes et complexes. Le *lieu* est défini dans le TLF en ligne[40] comme l'espace « déterminé par sa situation dans un ensemble, par la chose qui s'y trouve ou l'événement qui s'y produit » ou « qualifié par un adjectif qui le caractérise dans ses dimensions, son aspect, sa qualité ». Pour le Larousse en ligne[41], le *lieu* est une « situation spatiale de quelque chose, de quelqu'un permettant de le localiser, de déterminer une direction, une trajectoire » ou un « endroit, localité, édifice, local etc., considérés du point de vue de leur affectation ou de ce qui s'y passe ». Le Petit Robert (1993) définit le *lieu* comme une « portion déterminée de l'espace, considérée de façon générale et abstraite ».

Toutes ces définitions relient la notion de lieu avec celle d'espace, ou plus exactement une partie déterminée de l'espace, caractérisé par l'objet qui s'y trouve ou l'événement qui s'y passe. L'objet et l'événement peuvent être localisés, autrement dit, ils peuvent se positionner sur un plan ou une carte à l'aide de coordonnées géographiques.

Un autre terme proche de la notion de *lieu* est celui de *localité.* Il est défini dans

- le TLF comme « partie déterminée de l'espace » ;
- le Larousse comme « lieu déterminé constituant une entité géographique »
- le Petit Robert comme « lieu déterminé ».

Les deux mots *lieu* et *localité* se définissent presque de la même manière à partir de la notion d'*espace déterminé.*

Le dictionnaire Larousse emploie un nouveau terme (l'*entité géographique*) pour lequel les autres dictionnaires consultés ne proposent pas d'entrée.

La discipline qui étudie la désignation des lieux ou, comme il est dit dans la littérature spécialisée, le nommage des objets géographiques s'appelle la toponymie. Le dictionnaire de (Dubois *et al.* 1994) indique que la « toponymie est l'étude de l'origine des noms de lieux, de leurs rapports avec la langue du pays, les langues d'autres pays et des langues disparues. La matière est généralement divisée selon la géographie (il existe des spécialistes des noms de fleuves [hydronymie], des noms de montagnes [oronymie], des spécialistes pour telle ou telle région déterminée). » (p.485). La science de la toponymie distingue « des noms de lieux habités (villes, bourgs, villages, hameaux et écarts) ou non habités (lieux-dits) », « les noms liés au relief (oronymes), aux cours d'eau (hydronymes), aux voies de communication (odonymes, ou hodonymes) »[42], et des microtoponymes comme des noms de villas ou d'hôtels, par exemple. La toponymie semble ne pas établir de différence entre *toponyme* et *nom de lieu.* Les dictionnaires Larousse et Petit Robert vont dans le même sens en définissant le *toponyme* comme un « nom de lieu » et le TLF comme « nom de lieu de localité ».

Les documents officiels rédigés par les institutions spécialisées proposent leur propre définition. La Commission Nationale de Toponymie (CNT) indique qu'un toponyme dénote un objet géographique déterminé. On peut ajouter aux toponymes proprement dits des surnoms géographiques (*l'île de Beauté),* des noms géographiques « désignant des entités considérées du point de vue […] historique, culturel ou touristique » (*la Côte d'Or, la Côte d'Azur etc.*) ou provenant de « la coïncidence […] entre un territoire et une entité politique ou administrative » (*l'Ile de France, Neuilly-sur-Seine*). Pour l'IGN, « un toponyme est un nom de lieu, constitué d'un ou plusieurs mots, en rapport étroit avec un détail géographique localisé

[40] http://atilf.atilf.fr/dendien/scripts/tlfiv5/visusel.exe?12;s=2924586840;r=1;nat=;sol=1; [consulté le 10 mars 2014]
[41] http://www.larousse.fr/dictionnaires/francais/lieu/47076?q=lieu#47003 [consulté le 10 mars 2014]
[42] Définition d'un toponyme dans Wikipédia : http:///www.wikipedia.fr [consulté le 10/12/2012].

et avec le groupe humain qui l'utilise ». On distingue également les toponymes naturels des toponymes administratifs (pays, régions, villes etc.) désignant des espaces géographiques dont les limites ont été fixées par l'homme. La notion de *toponyme* est donc liée avec l'objet géographique qu'il dénote et sa typologie est déterminée par la nature de cet objet.

Les linguistes se sont intéressés à la notion du lieu à travers l'étude des expressions spatiales (Boons 1987, Borillo 1998, Laur 1991, Vandeloise 1986, Le Pesant 2011a, 2012 etc.). La thèse de Xavier Gouvert (2008) consacrée à la toponymie met en évidence la problématique et la complexité de la notion de lieu : « la tâche principale du lexicologue, du morphologiste et du syntacticien consiste à élaborer, justifier et/ou réfuter les critères définitoires de leurs catégories conceptuelles respectives, voire à en fonder de nouvelles. Il semblerait que le toponymiste puisse s'épargner cette peine : tout se passe comme si la notion de « nom de lieu » était une donnée directe de l'expérience ou une classe intuitive de la pensée » (p.137). L'auteur décrit également quelques propriétés linguistiques du fonctionnement des noms de lieu dans la phrase et constate qu'ils s'apparentent moins à un « nom propre » qu'à un « adverbe propre », dans la mesure où « les toponymes exercent primordialement et très majoritairement la fonction syntaxique de circonstant ou de second actant à signifié circonstanciel ».

Dans le domaine du traitement automatique des langues, les noms de lieu font partie des *entités nommées*. Les conventions de la campagne d'évaluation Ester[2] (2007)[43], par exemple, distinguent les lieux géographiques naturels, les régions administratives (*région Centre*, *Saint-Cyr-la-Rivière*), les axes de circulation (*autoroute A10, pont Gambetta),* les adresses et les constructions humaines (*prison de Fresnes, Musée d'Art Contemporain)*. La base de données des noms propres Prolex (Tran 1995) différencie les astronymes, édifices, géonymes, hydronymes, villes, voies etc. L'équivalence entre les catégories de Prolex et Ester a été décrite dans (Maurel *et al.* 2011). Dans les conventions Quaero (Rosset *et al.* 2011), les lieux sont désignés également comme des localisations et des entités spatiales. Sont distingués :

- les localisations administratives qui désignent « une portion de territoire dont le contour est géopolitique » (*Paris, La Maison Blanche, le 13^{e} arrondissement, la ville de Paris, la Bretagne, la France etc.*) ;
- les localisations géographiques : lieux physiques terrestres constituant des espaces géographiques naturels (*le désert de Gobi*) ; lieux physiques aquatiques ou hydronymes (*le canal Saint-Martin*) ; lieux physiques astronomiques (*la Lune*) ;
- les voies : oronymes (*l'autoroute A6*) ;
- les bâtiments et leurs extensions (*la gare de Rungis, l'Élysée*)
- les adresses : « un point […] dans l'espace » (*9 place de Rungis*) ; coordonnées électroniques (*mon numéro est le 01 69 85 80 02, mon identifiant skype est jean.dupont*) ;
- sites Internet : (*lemonde.fr*)
- plus quelques autres cas de moindre importance.

Un autre terme employé dans le TAL et qui évoque la notion de lieu est *entité spatiale.* (Lesbegueries 2007) propose la distinction entre une entité spatiale absolue caractérisant les informations propres à une entité nommée (*la ville de Paris*) et le concept d'entité spatiale relative caractérisant des indications spatiales associées aux entités nommées (*près de Paris*). Cette distinction est utilisée pour la recherche d'information dans les textes et dans la géolocalisation de cette information.

[43] http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf

La notion de *lieu* et son identification posent différents problèmes. En premier lieu, le nommage des objets géographiques n'est pas normalisé et vient souvent de la tradition orale. Deuxièmement, on ne peut pas limiter cette notion aux seuls noms propres car le lieu peut être désigné par les noms communs d'une manière neutre : *le village* ou personnalisée : *mon village.* Un endroit peut faire référence à un lieu imaginaire ou métaphorique *le bout de monde, mon paradis etc.* Les déictiques *ici, là* désignent aussi le lieu d'une manière référentielle. Une troisième difficulté provient des définitions proposées pour les toponymes qui s'appuient toutes sur la nature du référent désigné par le nom de lieu. Pourtant, un même lieu peut être désigné par plusieurs toponymes et inversement un même toponyme peut désigner des objets géographiques différents. (Maurel *et al.* 2011) note que « tous les toponymes sont susceptibles d'une interprétation comme anthroponyme collectif, toutes les associations ou entreprises peuvent être considérées dans certains contextes comme un toponyme ou un ergonyme ». Le même phénomène est mentionné par (Gouvert 2008) qui remarque que dans *Lyon est vaste* ou *Lyon bouge, Lyon* dénote plutôt l'objet matériel qui occupe ce lieu, c'est-à-dire une ville, et l'institution sociale qui s'y rattache, c'est-à-dire une commune. L'auteur distingue plusieurs sens des noms propres comme *Lyon* qui désigne à la fois (1) un lieu de la surface terrestre, (2) une agglomération urbaine et (3) une circonscription administrative et une communauté politique, ce qui conduit Gouvert à considérer le toponyme *Lyon* au sens (1) comme une unité linguistique différente de Lyon aux sens (2) et (3). Ensuite, la nature morphologique des mots désignant le lieu est variée. Ils peuvent être composés d'un ou de plusieurs mots, être liés ou non par un trait d'union. Du point de vue typographique, ils peuvent commencer ou non par une majuscule. En conclusion, la désignation d'un lieu est un processus social réapproprié subjectivement. L'emploi du nom de lieu dans le discours est une représentation de l'espace construite par le locuteur. La signification d'un toponyme, c'est-à-dire la référence à un lieu donné, est déterminée par le système toponymique du locuteur. Toutes ces difficultés rendent plus complexe la tâche d'identification automatique des lieux.

4.3. Corpus

Notre objectif a été d'analyser les désignations actuelles de lieu par les locuteurs variés. Pour cela, nous avons utilisé les ressources provenant du Web. Notre choix du corpus a été guidé par trois contraintes : la fréquence élevée des noms de lieu dans le corpus, la liberté de désignations de lieux et la multitude de locuteurs.

4.3.1. Constitution du corpus d'étude

L'Institut de l'information géographique et forestière (IGN) offre depuis 2007 un service web de « Carte à la carte » qui permet à tout utilisateur d'Internet de définir, à partir des bases de données géographiques de l'Institut, une carte topographique personnalisée en fonction de différentes caractéristiques : son format, son échelle, son orientation et son titre. La demande doit concerner une zone de France métropolitaine. L'usager centrant la carte sur un point, l'emprise de la carte est calculée à partir de ce centre. L'ensemble des demandes ainsi formulées constitue le corpus d'étude, 50942 titres :

> *normal;30000;paysage;626674;6829004;BIENVENUE A La Grosse Maison du Bouc Etourdi*

4.3.2. Nature du corpus étudié

Le corpus des titres de cartes est un corpus hors normes : il est composé des titres qui proviennent du Web.

4.3.2.1. Langage des titres

Le titre est un texte en tête d'un ouvrage, imprimé dans des caractères et une taille différents de ceux du corps de l'ouvrage. Souvent, les titres n'ont pas de signes de ponctuation et emploient des lettres majuscules et minuscules d'une manière peu conventionnelle. Le titre a une taille limitée et sert d'une manière générale à définir le sujet d'une œuvre et/ou attirer l'attention d'un futur utilisateur d'une œuvre. Cependant, les titres et leurs fonctions varient en fonction des ouvrages qu'ils représentent. (Hoek 2004) note qu'un titre de livre se propose plutôt d'informer sur son contenu et celui d'un film plutôt d'attirer l'attention du spectateur. Beaucoup de travaux ont été effectués sur les titres journalistiques. (Charaudeau 1983) démontre que « les titres, dans l'information, sont d'une importance capitale ; car, non seulement ils annoncent la nouvelle, non seulement ils conduisent à l'article (fonction « guide »), mais encore ils résument, ils condensent, voire ils figent la nouvelle au point de devenir l'essentiel de l'information. Le titre acquiert donc un statut autonome ; il devient un texte à lui seul, un texte qui est livré au regard des lecteurs et à l'écoute des auditeurs comme tenant le rôle principal sur la scène de l'information » (1983:102). Moirand (1975:69) parle également de « condenser en quelques mots le thème principal – « accrocheur » ou « illustrateur » – du message transmis par le texte ». Mouillaud (1982:84), à son tour, distingue entre les titres informationnels (qui résument l'article) et les titres référentiels (qui englobent l'article). Van Dijk (1985:69) souligne la fonction thématique du titre, celle d'exprimer le thème le plus important de l'article, tandis que les sous-titres au sein de l'article expriment les causes ou les conséquences importantes.

D'autres caractéristiques des titres journalistiques ont été soulignées par les chercheurs. Ainsi, (Engel 2000) a fait une étude comparative entre les titres de huit journaux, quatre français et quatre britanniques. Je présente ci-dessous quelques observations de l'auteur concernant la syntaxe des titres français :

- l'omission de la copule verbale : « Dans la presse française on note une forte préférence pour les titres sans verbe conjugué : de 60,6% à 78,6% » et l'omission de l'auxiliaire dans la phrase passive.
- pour les emplois verbaux, l'auteur note trois temps fréquents : le présent de 18,2% à 31,7%, le futur en deuxième position et enfin l'infinitif. Il relève également l'emploi de l'imparfait et du conditionnel ;
- l'omission de l'article dans le syntagme nominal ;
- la présence importante de nominalisations des syntagmes verbaux ;
- le point d'interrogation est l'un des rares signes de ponctuation possibles dans les titres.

En conclusion, le titre n'est pas une notion homogène. Il s'applique à des ouvrages de nature différente (article, livre, toile, film etc.). Ses fonctions varient : résumer, attirer l'attention du client, informer l'utilisateur sur le contenu de l'ouvrage, inciter le lecteur à acheter/lire/voir… cet ouvrage, aider à une meilleure compréhension. Le titre acquiert un statut autonome. Il peut constituer le corpus à lui seul et faire l'objet d'une étude spécifique (Engel 2000, Hoek 2004, Furet 1995, Mårdh 1980, Sullet-Nylander 1998, Lopez 2012).

4.3.2.2. Langage du Web

L'Internet a permis le développement de diverses formes de communication électroniques : courriel, liste de diffusion, messagerie instantanée, chat, blog, microblog, réseau social, sms.

Beaucoup de travaux en linguistique portent sur l'analyse des nouvelles formes de communication écrite (NFCE) en français. Les études sur la communication électronique ont été inaugurées dans les années 2000 par (Anis 1999, 2006). Les analyses du langage SMS

avec l'approche TAL ont été décrites dans (Véronis et Guimier De Neef 2006, Fairon *et al.* 2007, Fairon et Kein 2010, Panckhurst *et al.* 2013, Panckhurst et Moïse 2014). Diverses initiatives ont été prises pour constituer des corpus de NFCE : le projet « Faites don de vos SMS à la science » (Fairon *et al.* 2007), le projet « sms4sciences »[44] qui a produit le corpus « 88milSMS » ou encore le projet « CoMeRe »[45], lancé par le groupe TEI-CMC.

Parmi les travaux récents, on citera ceux de (Lorenz 2013) sur le langage du chat et (Longhi 2012, 2013) sur Tweeter. Ce microblog est au centre de beaucoup de recherches en sciences de l'information et de la communication (Honeycutt et Herring 2009, Java *et al.* 2007).

Les premières journées internationales de recherche (JIR) sur les Médias sociaux et les corpus de communication médiée par les réseaux (CMR) se tiendront à Rennes les 23-24 octobre 2015.

4.3.2.3. Corpus Web des titres de cartes géographiques

L'ensemble des titres de cartes forme un corpus textuel. Il est composé de 50942 lignes. Tous les titres ont une longueur limitée (55 caractères) et sont formés de phrases ou de groupes de mots servant à dénommer un objet : une carte géographique dans notre cas.

4.3.2.3.1. Analyse morphosyntaxique

Concernant la structure morphosyntaxique, les titres de cartes ressemblent aux titres classiques. Le Tableau 15 montre la distribution des catégories syntaxiques dans le corpus 2007 composés de 3387 titres.

Catégorie	Nombre total
Adjectifs	486
Adverbes	262
Conjonction de coordination	491
Déterminants	1689
Noms	10 782
Préposition	1358
Pronoms	145
Verbes	171

Tableau 15 : Distribution des catégories syntaxiques dans le corpus des titres

Une analyse statistique détaillée faite par Cordial est présentée dans sa totalité dans l'Annexe 16. Elle montre la distribution suivante par rapport à l'ensemble des mots:

- 26,00 % de mots-outils ;

Parmi les mots outils, on observe un emploi majoritaire des déterminants 12,52 % et la quasi-absence de conjonctions de subordinations 0,04 % ou de pronoms relatifs 0,06 % et possessifs

[44] http://www.sud4science.org/
[45] http://corpuscomere.wordpress.com/

0,00 % par rapport à l'ensemble des mots. Parmi les déterminants ce sont les articles définis qui sont employés majoritairement (8,95 %) par rapport à l'ensemble des mots.

- 74,00 % de mots signifiants (c.à-d. non mots-outils : substantifs, adjectifs, verbes, adverbes). Parmi lesquels : 92,15 % de substantifs, 4,15 % d'adjectifs, 1,46 % de verbes et 2,24 % d'adverbes.

Dans les pronoms, ce sont surtout les pronoms personnels qui sont utilisés :

- 23,42 % de pronoms personnels à la 1re personne du singulier
- 3,15 % de pronoms personnels à la 2e personne du singulier
- 36,49 % de pronoms personnels à la 3e personne du singulier
- 28,38 % de pronoms personnels à la 1re personne du pluriel
- 8,11 % de pronoms personnels à la 2e personne du pluriel
- 0,45 % de pronoms personnels à la 3e personne du pluriel

par rapport à l'ensemble des pronoms personnels.

Les catégories syntaxiques les plus fréquentes sont les substantifs (92,15 %), les déterminants (12,52 %) et les prépositions (8,66 %).

Ces résultats statistiques sur la distribution des catégories syntaxiques dans le corpus des titres de cartes ne sont pas surprenants et ils sont comparables aux autres corpus contenant les titres.

4.3.2.3.2. Caractéristiques particulières

Le corpus étudié provenant du Web, les conventions de nommage et d'écriture sont fluctuantes et varient d'un utilisateur à l'autre. L'expression peut être considérée comme proche parfois de l'oral :

Le Caylar En colombie... Si, si!

Pour maman de la part de tes quatre enfants

On y retrouve les caractéristiques principales des NFCE (Véronis et Guimier De Neef 2006):

- verlan

  A DONF

- graphies phonétisantes

  LADOUJEVIENS OUPETIGEAITAI

- squelettes consonantiques

  MASSIF MTB COL DE LA SEIGNE

- troncations

  JOYEUX ANNIF A la plus jolie des randonneuses

- étirements graphiques (saaalut ! biiiiizzzz !)

  CANTAL 07/13 Chouette alors !... Uuuuuuuuh....

- Didascalies et nouvelles conventions graphiques

  L'étoile :)

  CHEZ REGIS !!! Tours & Détours dans la Propriété CAILLE

Ce corpus, très hétérogène, a ses propres particularités. Les internautes ont des objectifs très différents dans la création et donc la dénomination de leur carte. Les titres ont des fonctions, des usages et des registres variés dans le corpus. Le titre peut être informatif et neutre comme

DEPARTEMENT DU VAR

VERCORS Presles (Isère) 26 Fév / 2 Mars 2013

ou subjectif et émotionné comme

VIVE YAYA! ...été 2013!

tip top chouette super genial youpi

On retrouve également l'humour ou le jeu de mots :

Ma Zone Avoir envie C'est être en vie

L'Emblavez à pied

Les thématiques des titres varient. Il peut s'agir d'un souvenir de vacances, de la préparation d'un événement partagé par une très petite communauté ou d'un cadeau à un proche. Les titres des cartes ont donc un caractére personnel et parfois intime. Cette nature des titres peut se manifester à travers l'intersubjectivité, l'expression du sentiment et/ou la présence des formes personnelles :

A NERAC Je me perds avec Hubert

Le Creux autour de chez nous

Les utilisateurs de « Carte à la carte » peuvent aussi appartenir à des communautés dont les activités s'appuient sur l'utilisation de cartes topographiques et utiliser le langage de ces communautés dans le titre de la carte, comme par exemple dans :

BLEAU TOP30

où *Bleau* est « l'appellation familière de la forêt de Fontainebleau dans les milieux sportifs, notamment le Groupe de *Bleau* »[46].

Tous les titres ne sont pas en français. Les internautes composent des titres en anglais, allemand etc., les mélangent ou utilisent des langues régionales comme le corse ou le basque :

U CAMPUTONDU CORSICA 2007

Le titre peut être complexe et contenir des sous-titres mais sans forcément avoir des marques de segmentation comme dans l'exemple suivant :

FAMILLE CHARRY LA TOUR BLANCHE 24 320 Perigord.

Les titres sont livrés sans contexte car on ne dispose pas de renseignements sur l'utilisateur lui-même, ni sur les motifs de sa demande. L'analyse du corpus se trouve restreinte à ce corpus composé d'une information sous forme textuelle (le titre même) et numérique (les coordonnées géographiques de l'endroit sélectionné).

Les spécificités du corpus décrites ci-dessus multiplient le nombre d'interprétations possibles et rendent son traitement automatique difficile. Voici les objectifs que nous avons fixés :

- le repérage et l'annotation des noms de lieux en nous fondant sur la nature du corpus,

- l'analyse des variations dans la désignation de lieux par les différents utilisateurs,

[46] Wikipédia : http://fr.wikipedia.org/wiki/Groupe_de_Bleau [consulté le 12/12/2012]

- l'étude de l'écriture de ces lieux sur le Web,

- l'analyse de la perception des lieux et des sentiments qu'ils évoquent chez l'utilisateur.

4.4. Noms de lieux dans le corpus

La partie qui suit est consacrée à l'observation des noms de lieux présents dans le corpus. Deux phénomènes se sont imposés à notre attention : le non respect des règles d'écriture des noms de lieux, source de multiples variations et la présence des lieux appelés *subjectifs.* Ces observations ont déterminé les modalités de leur annotation automatique.

4.4.1. Ecriture des noms de lieu

4.4.1.1. Règles d'écriture

L'écriture des toponymes diffère selon l'usage : les panneaux indicateurs, les plaques indicatrices de rue, le Web. Les règles d'écriture des toponymes existent, mais elles sont compliquées, subtiles et non homogènes, d'où des difficultés de compréhension et de mise en application. Deux signes typographiques en particulier rendent difficile l'écriture de toponymes composés : la majuscule et le trait d'union.

4.4.1.1.1. Emploi de la majuscule

L'usage problématique de la majuscule dans les noms propres a été relevé par les linguistes. Ainsi, (Mathieu-Colas 1998) note l'absence de cohérence et la variabilité des règles d'emploi. Parlant des dictionnaires et d'ouvrages de référence, il souligne que « tous s'attachent à décrire, avec beaucoup de soin, les mille et un secrets des majuscules […] Si chaque auteur présente ses règles sous une forme impérative, on note de l'un à l'autre un certain nombre de divergences qui, dissipant l'illusion d'une norme universelle, ne font que mettre en évidence l'instabilité du système ». Où et quand mettre une majuscule ? Comment vérifier le bon emploi d'une majuscule ? Sur quel ouvrage se fonder ?

Pour essayer de répondre à ces questions, j'ai consulté trois ouvrages : le *Bon Usage* de (Grevisse et Goosse 1993), les recommandations et observations grammaticales de la CNT (Commission nationale de toponymie) et la thèse de Mounira Bioud (2006). (Grevisse et Goosse 1993) consacrent un chapitre à l'emploi des majuscules et plus particulièrement au cas des noms de lieux (p.107-108) qui sont pour lui les « villes, villages, régions, pays, îles, montagnes, cours d'eau, mers […] étoiles, astres […] noms de rue […] noms des points cardinaux ». Les auteurs constatent l'usage flottant de la majuscule dans plusieurs cas :

- les noms des points cardinaux employés avec un complément qui est lui-même un nom de lieu : *dans le nord de la France* / *dans l'Ouest de la France* ;
- un nom de lieu désignant un objet : *un morceau de Brie / un morceau de brie.*

Pour les noms de lieux composés d'un nom propre et d'un nom commun, le nom commun ne prend pas de majuscule (*l'île Maurice, la ville de Paris*) sauf s'il « fait partie intégrante du nom propre » (*Val-d'Isère, le Val d'Aoste*).

Les recommandations de la CNT constituent une sorte de « bon usage » à destination des rédacteurs de toponymie et des collectivités locales. Elles indiquent, en premier lieu, que la question de la majuscule ne se pose pas dans les cas d'une inscription toponymique sur une carte ou sur un panneau indicateur car cette lettre y est toujours présente comme en début de phrase et dans les cas d'un toponyme simple, composé d'un seul mot « signifiant » (*Paris, la Seine*). Dans un toponyme composé, prennent la majuscule :

- « les mots significatifs » (substantifs, adjectifs, verbes ou adverbes) s'ils font partie d'un groupe de mots joints par un trait d'union (*les Champs-Elysées*) ou ayant une fonction de complément avec ou sans préposition (*la côte d'Or, la ville Lumière*) ;
- les substantifs employés en tant que noms propres, c'est-à-dire « dans un autre sens que leur sens habituel » (*le Val de Loire, le Crêt de la Neige*) ;
- « les adjectifs modifiant le sens des termes qu'ils qualifient » (*le mont Blanc, la côte Vermeille, la ville Éternelle*) et « les substantifs ainsi qualifiés s'ils sont placés après ces adjectifs (*l'Ancien et le Nouveau Monde*) ;
- l'article initial non contracté (*La Rochelle, Le Puy*).

La thèse de Mounira Bioud (2006) traite de la normalisation de l'emploi des majuscules dans les noms propres pour établir un système de vérification orthographique. Selon l'auteure, la règle générale est « de mettre une majuscule au spécifique et une minuscule au générique » (*la mer Rouge, le mont Everest, l'île d'Yeu* et de mettre une majuscule aux deux mots s'ils « forment une entité inséparable nécessaire à l'identification » (*le Pays Basque, Terre-Neuve, Golf-Juan, Val-d'Isère*). Les cas particuliers sont :

- si l'adjectif marque l'appartenance, il reste en minuscule (*la Gaule cisalpine*) ;
- si l'adjectif indique une position géographique, il reste en minuscule (*la haute Garonne*)
- si le toponyme a une forme *Adj de Npr*, l'adjectif et le nom propre commencent par une majuscule (*les Hauts-de-Seine*[47])
- pour les cas où le toponyme est précédé par un article, celui-ci prend une majuscule si le nom désigne une commune (*Le Havre, Le Mans*).

Le tableau (Tableau 16) dresse une synthèse de ces trois ouvrages concernant les cas de deux toponymes *le mont-Blanc* et *le massif du Mont-Blanc.* Pour la même écriture, les trois travaux proposent des explications différentes. En premier lieu, les auteurs n'utilisent pas les mêmes termes. (Grevisse et Goosse 1993) distinguent *le nom propre* et *le nom commun*, (Bioud 2006) différencie *générique* (nom commun) et *spécifique* (adjectif) alors que les recommandations de la CNT séparent les *mots significatifs* des *non significatifs*. En deuxième lieu, les règles sont complexes et non homogènes. Elles manquent aussi d'unité et d'harmonisation. Enfin, elles supposent des connaissances syntaxiques préalables.

Certaines règles sont extrêmement floues ; en particulier, dans quel cas est-il possible de déterminer si le nom commun fait partie ou non du nom propre, le trait d'union n'étant pas toujours présent ? Les difficultés pour généraliser les règles sont réelles.

4.4.1.1.2. Utilisation du trait d'union

Le trait d'union est un des critères de l'emploi de la majuscule dans les toponymes composés. Pourtant les règles de son emploi sont aussi compliquées et fluctuantes que pour la majuscule. La CNT, par exemple, juxtapose des critères sémantiques et syntaxiques : « Parmi les mots composant en français un toponyme […] sont joints par des traits d'union les mots ayant perdu dans la composition leur sens ou leur syntaxe habituels ». L'un des sous-exemples de cette affirmation concerne « les mots appartenant à un groupe de mots ayant une fonction de complément (avec ou sans préposition) au sein du syntagme toponymique et ne se limitant pas à décrire l'objet géographique » : *le massif du Mont-Blanc*, *le parc des Buttes-Chaumont*. Cependant lorsque ces mots n'ont pas la fonction de complément, cette règle n'est plus valable (*le mont Blanc*) et la majuscule du nom générique est enlevé. (Grevisse et Goosse 1993), traitant les signes typographiques dont fait partie le trait d'union et sans distinguer le cas des toponymes, indiquent la présence de ce signe « à la suite d'un changement de

[47] Cet exemple donné par Mounira Bioud me semble peu heureux car *haut* ici est un nom.

signification » (*la rue Saint-Pierre, la ville de Saint-Etienne*). Pour (Bioud 2006), « les noms propres reliés par un trait d'union forment une suite figée », donc inséparable, et prennent tous deux une majuscule initiale. Comment le scripteur pourrait-il mémoriser des nuances typographiques qui présupposent des connaissances linguistiques approfondies, requérant une analyse syntaxique et sémantique préalable du syntagme écrit et la connaissance du sens et de la structure qui a fait l'objet d'une transformation.

	le mont-Blanc	***le massif du Mont-Blanc***
BU	l'adjectif prend la MAJ quand il accompagne un nom commun géographique	en position de complément, Mont prend la majuscule et se lie au mot suivant (*la beauté du Mont-Blanc*)
CNT	les adjectifs modifiant le sens des termes qu'ils qualifient avec une précision suffisante pour désigner un nouveau toponyme prennent la MAJ	les mots signifiants, appartenant au sein du syntagme toponymique à un groupe de mots ayant une fonction de complément prennent une majuscule
Bioud	si le nom est un générique, alors le nom générique reste en min, l'adjectif spécifique commence par une MAJ	si le premier nom est générique, alors il reste en min. et les composants du spécifique commencent par une MAJ et sont reliés parun trait d'union

Tableau 16 : Synthèse des règles d'emploi de la majuscule dans les toponymes sur un exemple concret

Il n'existe donc pas de consensus réel entre les auteurs concernant l'usage de la majuscule ou du tiret dans la plupart de leurs emplois, ce qui encourage des pratiques disparates dans la notation des toponymes : « Qu'on opte pour une « harmonisation orthographique » […] ou pour une tolérance bien tempérée, il convient de se libérer des « délires » de l'orthographe […]. Rien ne serait pire pour le traitement automatique que de vouloir s'accrocher à des normes aussi pointilleuses qu'arbitraires. » (Mathieu-Colas 1998:12). Cette absence d'unification des règles d'usage laisse le scripteur plus libre de ses choix graphiques. La méthode proposée pour identifier les noms de lieu dans le corpus tient compte de cette liberté orthographique. L'analyse des variations d'écriture montre un accroissement des idiosyncrasies, signalant une orientation nouvelle dans l'emploi scriptural de ces noms.

4.4.1.2. Analyse de l'écriture des noms de lieu dans le corpus

Pour rendre sensible cette variabilité, observons l'exemple suivant où le même lieu *Monts d'Arrée* est orthographié de sept manières différentes dans le corpus :

> *MONTS D ARREES / MONTS D ARREE / MONT ARREE / Monts D'arrée / monts d'arrée / monts d arree / Monts d Arrée*

Les variations concernent l'emploi de la majuscule, de l'apostrophe, de la marque de pluriel *s* et de la préposition *de*[48].

Le corpus étant issu du Web, l'analyse des noms de lieux dans le corpus des titres de carte a permis d'observer les diverses NFCE. Les quatre phénomènes caractéristiques de l'écriture des noms de lieux dans le corpus ont été déterminés rapidement :

[48] Notons que dans la variation terminologique, de telles variations sont aussi très fréquentes.

- *troncations*
- *agglutinations*
- *néologismes*
- *contact de langues*
- *erreurs de frappe*

4.4.1.2.1. Troncation

On peut voir la troncation comme l'un des mécanismes les plus radicaux de la créativité lexicale. Les cas de troncation sont assez courants surtout dans le langage familier écrit, oral ou sur le Web comme par exeple *cata* (pour *catastrophe*), *diapo* (pour *diapositive*), *appli* (pour *application*) etc. Nous avons observé le même phénomène chez les noms de lieux.

Si le lieu est désigné par un seul mot, la troncation peut porter :

- sur la fin du mot sans modifications comme dans *Monac (*pour *Monaco), ch (*pour *chemin), agglo (*pour *agglomération)* – ou avec modification comme *Manox* (pour *Manosque*), *Bézo* (pour *Bézaudun*),
- sur le centre du mot *Dne (*pour *domaine).*

Si le lieu est représenté par un nom composé, la troncation peut porter :

- sur le dernier mot : en ne gardant que la première lettre *Vitry sur S.* (pour *Vitry-sur-Seine*), *Alpes M.* (pour *Alpes-Maritimes*), en le tronquant partiellement *La Tour d'Auv* (pour *La Tour d'Auvergne*) et en le supprimant définitivement *aix*, *CHAMPAGNAC_LA* (pour *Champagnac-la-Rivière*).
- sur tous les mots : création d'un sigle *SQY (*pour *Saint-Quentin en Yvelines)*, réduction au squelette consonantique *MBT* (pour *Mont-Blanc*), *NTS ST GEORGES* (pour *Nuits-Saint-Georges*).

4.4.1.2.2. Agglutination

Le phénomène d'agglutination est bien représenté dans le corpus. Il concerne surtout les noms de lieu composés et consiste dans :

- la suppression de séparateurs (blanc, tiret, apostrophe) :

  LADOUJEVIENS, BoisHébert, BortLesOrgues ;

suivie parfois de la troncation

  *BgArgental (*pour *Bourg-Argental), ISSOULET Etapapy 2007 (*pour *Etape à papy)*

- la suppression des mots vides (articles, prépositions) :

  *CheminStevenson (*pour *chemin de Stevenson)*

  *CinqueuxSenlis (*pour *Cinqueux fait partie de Senlis)*

Les deux phénomènes de troncation et d'agglutination sont liés probablement à la limitation de longueur du titre (55 caractères). L'utilisateur raccourcit le nom de lieu en considérant que l'information est connue du ou des destinataire(s) de la carte. Le fait de le modifier en changeant son écriture ne nuira donc pas à sa compréhension. Une autre raison possible est que ces « nouvelles écritures » peuvent être tout-à-fait familières à la communauté à laquelle s'adresse le demandeur de la carte.

4.4.1.2.3. Néologismes

L'analyse de l'écriture des noms de lieu montre une présence importante de néologismes.

Nous avons reconnu nombre de créations utilisant le suffixe germanique *land.* Cette création est très productive dans notre corpus :

GRIDOULAND, Tamalou-land, Plouc land

L'ajout d'un suffixe à un nom ou surnom se fait avec ou sans séparateur.

La création lexicale dans le corpus peut être singulière comme dans

L' ânexxe SERILLY

où l'utilisateur fait une crase d'*âne* et *annexe* par leur combinaison graphique, utilisant l'agglutination et le doublement de la consonne.

Les néologismes liés aux lieux peuvent être créés aussi par le remplacement d'un mot par un autre :

*CORDES EN CIEL (*pour *CORDES SUR CIEL)*

En remplaçant la préposition *sur* par *en,* l'utilisateur semble faire un jeu de mots avec *l'arc-en-ciel.* Le lieu n'a donc plus une connotation neutre mais il est perçu d'une certaine manière par l'utilisateur.

Un autre procédé de création remarqué est l'assonance entre deux mots désignant le lieu :

AZZAZ et ses environz (pour *environs*)

Le remplacement du *s* final des *environs* par un *z* est probablemet lié à l'intention de l'utilisateur de faire apparaître une assonance avec *Azzaz*. En outre, le son *z* peut faire allusion au bruit ce qui induit une certaine perception et une interprétation personnelle du lieu par le demandeur.

4.4.1.2.4. Contact des langues

L'utilisateur mélange parfois plusieurs langues pour désigner un même lieu :

CHEZ LAURENT and surrounds, El Carlit

Comme on voit dans ces exemples, l'usage d'une langue étrangère porte sur l'information circonstancielle ou sur le déterminant, l'indication du lieu même restant en français. Nous avons vu dans la section précédente que l'influence d'une langue étrangère se manifeste également à travers l'invention de noms de lieu par le suffixe *land.*

4.4.1.2.5. Erreurs de frappe

Les erreurs de frappe sont très fréquentes dans le corpus, ce qui est tout à fait représentatif du corpus Web.

LA MONTAGHE (pour *LA MONTAGNE*)

lA bARROTI7RE (pour *La Barrotière*)

4.4.1.2.6. Autres phénomènes

Parmi d'autres phénomènes caractéristiques des NFCE moins fréquents dans le corpus, mentionnons :

- les graphies phonétisantes ou des transcriptions phonétiques :

NOT'BARAQUE ché par ichi ;

LADOUJEVIENS OUPETIGEAITAI

- les émoticons et les nouvelles conventions graphiques :

Le TERRAIN HI HI HI

Gare au Loup :0) :0) :0)

Charlat et ses alentours !!

Cette variété et créativité de désignations de lieux dans le corpus a permis de dégager un phénomène supplémentaire. La désignation d'un lieu n'est pas toujours neutre. Nous avons appelé les lieux ainsi désignés des *lieux subjectifs*.

4.4.2. Lieux subjectifs

La notion de la subjectivité est aujourd'hui au centre de nombreuses recherches en TAL, en lexicographie, en linguistique de corpus, et en analyse de discours. En effet, Internet et les nouvelles technologies ont permis une nouvelle forme de communication à travers les réseaux sociaux, blogs, sites commerciaux etc. caractérisée entre autres par une expression massive de sentiments et d'opinions.

Les études concernent en premier la constitution de ressources lexicales spécifiques : les adjectifs et les groupes adjectivaux (Hatzivassiloglou et McKeown 1997), les verbes (Vanderveken 1988), les adverbes évaluatifs (Guimier 2002), les prédicats de sentiments (Mathieu 1999), les mots d'affects (Le Pesant 2011b, Bradley et Lang 1999). Pour une approche plus informatique, je citerai les travaux d'(Esuli et Sebastiani 2006) décrivant une ressource lexicale SentiWordNet (Baccianella *et al.* 2010) qui associe deux propriétés supplémentaires : la subjectivité et la polarité aux synsets de WordNet (Fellbaum 2005). Quand le TAL vise la détection automatique des opinions et des sentiments, l'analyse de la subjectivité se contente souvent de l'attribution d'une polarité (positif, négatif ou neutre). Au plan méthodologique, ce sont les techniques statistiques de la fouille des données qui semblent dominer la recherche (Pak et Paroubek 2010, Marchand *et al.* 2014). Des chercheurs proposent aussi des outils fondés sur l'analyse linguistique. C'est le cas de l'outil FMO (Bouraoui et Canitrot 2013) qui permet d'extraire du texte des segments porteurs d'opinions pour leur attribuer ensuite une note selon la polarité. (Vernier 2011) propose une méthode automatique pour délimiter les passages subjectifs dans un corpus de blogs multidomaines, puis pour les catégoriser selon leur modalité et leur polarité. Dans l'analyse du discours, les travaux de (Kerbrat-Orecchioni 1999) ou ceux de (Plantin 2011) font référence. Différents projets ont mis au centre de leurs préoccupations la subjectivité langagière : par exemple EMOLEX[49] qui « vise à analyser les valeurs sémantiques, le comportement combinatoire (lexématique et syntaxique) et les profils discursifs des lexies des émotions dans cinq langues européennes » et qui a donné naissance à une base de données EmoBase[50] ; ou encore OntOpiTex[51] qui cherche à identifier, agréger et classer des segments textuels porteurs d'opinions. Enfin, un autre projet, Senterritoire[52], a comme objectif de détecter les opinions et les sentiments liés à l'aménagement d'un territoire.

[49] http://www.emolex.eu/
[50] http://emolex.u-grenoble3.fr/emoBase/
[51] https://ontopitex.greyc.fr/
[52] http://www.msh-m.fr/la-recherche/programmes-actuels/senterritoire/

Le travail sur les désignations de lieux dans le corpus des titres de carte a montré que la subjectivité peut être aussi liée à des notions géographiques.

4.4.2.1. Appropriation, personnalisation de lieu :

Le lieu peut être approprié et de cette manière personnalisé par le locuteur. Les structures syntaxiques sont caractérisées dans ce cas par la présence de :

- déterminants ou pronoms possessifs accompagnant le lieu en question :

  ma campagne, Ma Montagne Saint Sorlin, ma zone à moi, notre fief, notre coin

- complément humain introduit par les prépositions *de* ou *à* et ayant comme fonction de désigner le possesseur d'un lieu *:*

  fief de Patrick XXX, coin de Guy, village de ma Maman

- prénom ou nom de personne se trouvant dans un contexte propre aux lieux (après les prépositions ou adverbes locatifs) :

  Autour de Sylvie et Dom, environs de Edith et Luc

- création lexicale formée d'un mot suffixé par -*land* ajouté au nom désignant un humain

  Papyland, Zozo land, TAMALOU-LAND, PLOUCLAND

4.4.2.2. Lieux imaginaires

Le lieu peut appartenir au monde des œuvres artistiques :

Les Terres du Milieu, Atlantide

4.4.2.3. Lieux métaphoriques

Le lieu peut être désigné en comparaison avec d'autres lieux existants. Il s'agit des lieux en France comparés ou associés à un lieu étranger :

Amérique Française

4.4.2.4. Evaluation du lieu et/ou avec expression d'un sentiment

Le lieu peut être évalué par le locuteur et comporter des marques de jugements. Je distingue plusieurs cas possibles :

- lieu est accompagné par ce qu'on appelle du « lexique de sentiment ou d'appréciation » :

  lieu magique, pays préféré, tanière idéale, villa aimée, bons coins, joli petit coin, ILE des rêves, maison du bonheur

- unités lexicales utilisées pour dénommer le lieu contenant déjà une évaluation ou un sentiment :

  trou du cul, centre du monde, bout de chemin, sweet home

  paradis, berceau

- présence de signes typographiques :

  Le Caylar En colombie... Si, si!. la maison de ... Guy XXX, maison :), "Nice Aire St Michel"

- jeux de mots, imitations phonétiques ou comparaison/association du lieu aux autres sujets :

NOT'BARAQUE ché par ichi, CORDES EN CIEL, AZZAZ et ses environz, L' ânexxe SERILLY

Ces derniers exemples montrent les cas difficiles de prendre en compte dans le processus de repérage automatique des lieux. Il s'agit de processus singuliers propres aux utilisateurs et donc irréguliers.

4.4.2.5. Lieux accompagné d'une fonction

Le lieu peut être accompagné par une mention de fonction (introduite souvent par un groupe prépositionnel en *à*) :

chemins à parcourir, circuits à fleurs et à champignons, chemins à découvrir, à explorer

Comme on peut le voir dans les cas présentés ci-dessus, la subjectivité dépend du contexte de l'utilisation d'un lieu. Le lieu devient subjectif ou le nom devient un lieu grâce au contexte linguistique :

- la présence de déterminants ou pronoms possessifs ou du lexique de sentiments

  notre fief, lieu magique

- le nom qui ne désigne pas habituellement un lieu est employé dans le contexte propre au lieu (derrière les prépositions ou adverbes locatifs)

  randonner autour de notre AMOUR, balades autour de Germaine, Paris et Pierre

- la coordination entre deux noms de lieu de niveaux différents :

  THOMERY ou chez Pierre et Marie

et extralinguistique : l'emploi du lieu étranger dans le corpus des titres de cartes[53]

Afrique du sud

L'analyse des désignations de lieux dans le corpus de titres a guidé notre travail sur leur repérage automatique. Nous avons essayé de tenir compte d'une part de leur variation orthographique et d'autre part de l'aspect subjectif dans leur désignation.

4.5. Repérage automatique

Les noms de lieux font partie des entités nommées ; leur reconnaissance automatique est étudiée dans différents travaux en traitement automatique de la langue (Maurel *et al.* 2011, Bouamour 2009, Gaio *et al.* 2012) qui proposent des méthodes symboliques, probabilistes ou hybrides.

Le lieu peut être désigné directement par le toponyme. Dans ces cas, pour le reconnaitre automatiquement, on peut utiliser les ressources lexicales recensant les noms de lieux (noms propres) de la France. Pourtant, on ne peut pas se contenter du repérage des toponymes car les lieux peuvent être désignés avec les noms communs employés dans le contexte lexical ou syntaxique des lieux. Pour reconnaître ces derniers, on va se servir de ce contexte. La technique la plus appropriée est la méthode symbolique utilisant les grammaires locales. La méthode proposée s'appuie ainsi, d'une part, sur l'utilisation des ressources lexicales

[53] Le corpus traité est constitué à partir d'un service web de « Carte à la carte » qui permet à tout utilisateur d'Internet de définir, à partir des bases de données géographiques de l'IGN, une carte topographique personnalisée. La demande ne doit concerner qu'une zone de France métropolitaine.

contenant les toponymes français et d'autre part, sur la description du contexte de l'emploi du nom de lieu sous forme de patrons.

4.5.1. Ressources existantes pour identifier des toponymes français

Les ressources lexicales contenant les noms de lieux sont nombreuses, variées et visent des objectifs différents.

Wikipédia constitue une ressource riche et en perpétuelle incrémentation. Cependant des difficultés d'utilisation apparaissent à cause de la diversité des contributeurs et par conséquent de l'hétérogénéité des définitions proposées ce qui rend difficile la comparaison entre ces entrées. Une autre ressource existant sur le Web est Geonames, une base de données de lieux.

D'autres ressources numériques comme la base de données des noms propres *Prolex* (Tran et Maurel 2006) qui contient, entre autres, les toponymes ou non numériques comme *le Dictionnaire étymologique des noms de lieux en France* (Dauzet 1963) permettent de définir des listes de toponymes.

L'IGN a constitué une ressource lexicale spécifique recensant les toponymes de France métropolitaine, BDNyme. Sa couverture représente plus de 1,7 million d'entrées. C'est une ressource de référence pour la reconnaissance des toponymes de la France métropolitaine.

4.5.2. Méthode employée

Pour repérer les désignations des lieux dans les titres, deux méthodes ont été utilisées :

- la première a pour objectif de reconnaître les toponymes sans contexte en utilisant la base BDNyme de l'IGN (ce travail a été fait par Catherine Domingues) ;
- la deuxième procède à une analyse de surface à l'aide de la plateforme Unitex (Paumier 2003) et utilise des grammaires locales pour repérer les noms de lieux et d'autres expressions spatiales dans le corpus (cette deuxième partie a été effectuée par moi).

Le repérage des désignations de lieux se décompose en plusieurs étapes successives. A chaque étape, les toponymes reconnus dans le corpus sont balisés et typés. Le résultat de ces transformations constitue le corpus d'entrée de l'étape suivante. La méthodologie proposée s'appuie sur le contexte et le respect de la nature des données. Ce recours au contexte recouvre deux aspects : le géoréférencement de la carte (l'emprise exacte : étape 1 et l'emprise élargie : étape 2) et la description du contexte linguistique à l'aide des patrons (étape 3). La méthodologie est guidée par l'analyse préalable du corpus avec la prise en compte de ses caractéristiques : la variation orthographique et la nature diverses des lieux mentionnés. L'extrait du corpus annoté est présenté dans l'Annexe 15.

4.5.2.1. Utilisation de BDNyme

La base de données de l'IGN, BDNyme, ayant été utilisée par Catherine Domingues, on se contentera d'en donner un bref aperçu.

L'identification des toponymes repose sur la comparaison des chaînes de caractères de BDNyme et celles contenues dans les titres. L'utilisation de BDNyme se fait en deux étapes :

Etape 1 : ne sont examinés et recherchés dans le titres que les toponymes qui sont situés dans l'emprise exacte de la carte[54] ;

[54] Cette localisation est calculée grâce aux coordonnées géographiques du toponyme et à celles du centre de la carte désigné lors de la demande

Etape 2 : l'emprise est élargie en incluant la prise en compte des variations orthographiques comme :

- l'absence ou la présence de majuscules :

  LILLE , Lille, lille ;

- l'absence ou la présence de signes diacritiques :

  FRESNES-LES-MONTAUBAN et FRESNES-LÈS-MONTAUBAN ;

- la variation entre les séparateurs : blanc, trait d'union ou apostrophe ;
- l'acceptation des abréviations comme *st* et *ste* pour *saint* et *sainte* ;
- l'omission des mots vides : déterminants, prépositions ;
- l'abréviation d'un toponyme composé à condition que les mots du toponyme ne soient pas un mot vide ou l'adjectif saint(e), un générique de noms de lieux (plage, champ etc.), ni un prénom. Par exemple, dans le titre :

  AUTOUR DE BOUC Attention ça grimpe,

  BOUC est reconnu comme le nom abrégé de *Bouc-Bel-Air*.

Le fichier des titres est ainsi balisé :

{CHOLET,.<ChefLieuE2>} {Forêt de Nuaillé,.<ToponymeDiversE1>}

Dans cet exemple, *Forêt de Nuaillé* est reconnu comme un toponyme dans l'étape 1 et *Cholet* comme un chef-lieu dans l'étape 2.

4.5.2.2. Méthode symbolique

La deuxième étape, complémentaire à la précédente, est le repérage des toponymes en contexte à l'aide de patrons.

La typologie des noms de lieu a été développée, elle comprend 4 classes :

LieuGénérique : il s'agit des lieux formés à partir d'un nom commun définissant le type de ce lieu, qu'ils soient employés seuls ou accompagnés d'un complément ou d'un qualifiant.

Cette classe est composée à son tour de 17 sous-classes :

- LieuBatimentAnimal : *écuries, terrier, tanière, poulailler etc.*
- LieuBatimentHum : *chalet, maison, villa, résidence etc.*
- LieuChemin : *chemins, pont, tunnel etc.*
- LieuCommercial : *magasin, boutique, marché etc.*
- LieuEducatif : *école, académie, lycée etc.*
- LieuEntreprise : *entreprise, coopérative, firme etc.*
- LieuEtablissement : *agence, office, institution etc.*
- LieuHistorique : *fontaine, forteresse, château, castel, statue, tombeau etc.*
- LieuHotel : *hôtel, chambre d'hôtes, gîte, palace, auberge etc.*
- LieuMedical : *hôpital, clinique, laboratoire etc.*
- LieuNaturel : *forêt, île, lac, montagne, mer, marais etc.*
- LieuPrison : *prison, police, préfecture etc.*
- LieuReligieux : *église, basilique, chapelle etc.*
- LieuRestaurant : *brasserie, café, pizzeria etc.*
- LieuSportif : *stade, piscine, base etc.*
- LieuTerritoire : *secteur, agglomération, banlieue, capitale, cité etc.*

- LieuTransport : *gare, aéroport, port etc.*

LieuSubj désigne les lieux appropriés et personnalisés par l'utilisateur (*mon paradis, far east)* et/ou les lieux imaginaires comme (*Tamalou-Land).* Les règles de reconnaissance sont fondées sur le contexte particulier d'un lieu générique qui devient subjectif avec le pronom possessif, un complément humain ou un adjectif d'appréciation, sur le suffixe *land* ou sur le lexique établi (*paradis, bout du monde etc.*).

LieuDeict est utilisé pour les déictiques (*là-bas, ici etc.*)

LieuAdresse désigne une adresse.

L'évaluation de la méthode de la reconnaissance automatique des lieux dans le corpus des titres de cartes est décrite dans (Dominguès et Eshkol-Taravella 2013, 2015, à paraître).

4.6. Bilan

4.6.1. Compte rendu

Objectif visé	Etudier les désignations actuelles de lieu par des locuteurs variés. Pour cela, repérer et annoter l'information spatiale dans un corpus de titres de cartes issu du Web.
Méthodologie	Méthode symbolique utilisant les ressources lexicales extérieures et les grammaires locales construites.
Données traitées	Corpus des titres de cartes géographiques issu du service Web « Carte à la carte » proposé par l'IGN
Difficultés rencontrées	Des caractéristiques propres aux corpus provenant du Web : variations d'écriture, troncation, agglutination, néologismes, erreurs de frappe etc. Des particularités du corpus traité : contact des langues, absence de la segmentation entre titre et sous-titre, manque de contexte, caractère intime et personnel
Résultats	Corpus : les 50942 titres de cartes annotées en désignations de lieux Analyse des variations dans la désignation de lieux Etude des variations d'écriture de ces lieux sur le Web Etude de la perception des lieux chez l'utilisateur Elaboration d'une typologie des lieux *subjectifs*
Contraintes et limites	Certains noms de lieu n'ont pas été annotés à cause de leur irrégularité
Originalité et apport du travail	Traitement automatique adapté au corpus hors normes très particulier et unique dans son genre Méthode proposée fondée sur la nature des données et sur les variations linguistiques qui y sont présentes Etude de la subjectivité à travers les noms de lieu (les travaux antérieurs sur les expressions locatives, sur les entités nommées ou sur les entités spatiales d'une part et sur la subjectivité dans le langage d'autre part n'ont pas mentionné ce phénomène alors qu'il est présent ailleurs que dans notre corpus)

4.6.2. Perspectives et travaux en cours

Les lieux subjectifs restent pour moi un objet d'étude. Nous envisageons une annotation plus fine afin d'effectuer des études quantitatives sur l'emploi en discours. Cette annotation permettra de mieux observer la variété des procédés mis en œuvre par un locuteur pour désigner d'une manière subjective un endroit. L'annotation manuelle des lieux subjectifs sera effectuée par deux annotateurs afin de calculer l'accord inter-annotateur de la restitution de ce phénomène (les conventions de cette annotation sont présentées dans l'Annexe 17).

Nous envisageons de tester une autre méthode : l'apprentissage automatique. Pour cela, l'annotation manuelle de tous les lieux sera effectuée avec comme objectif de créer un corpus de référence.

Nous continuerons aussi le traitement du corpus des titres en procédant à une annotation de toutes les informations ce qui permettrait de mieux cerner la demande de cartes des usagers de ce service : les destinataires (*la carte pour papi*), les encouragements (*en avant*), les éléments temporels (*été 2007*), les événements (*20 ans de mariage*). Une des finalités serait d'adapter les typographies, les légendes de cartes, les illustrations de couverture etc. aux demandes des usagers des services cartographiques disponibles sur le Web. Cette perspective s'inscrit dans le cadre plus large de la recherche et de l'exploitation d'informations spatiales contenues dans du texte.

Le travail du repérage de l'information spatiale dans le corpus des titres peut être appliqué aux autres corpus. Le test sur le corpus oral ESLO sera effectué pour pouvoir y repérer les désignations des lieux et les sentiments que ces lieux évoquent pour les Orléanais.

5. Etude des noms généraux dans le corpus médiatique.

Cette partie concerne le corpus écrit médiatique. Elle est consacrée à l'étude de noms généraux fréquents dans ce type de corpus. Même si le corpus médiatique fait partie des corpus dit normalisés, l'emploi des noms généraux y offre certaines spécificités. Cette recherche a été conduite en collaboration avec Silvia Adler, enseignante de sémantique à l'Université Bar-Ilan.

5.1. Résumé du travail

Les noms généraux sont au centre de recherches en linguistique et linguistique de corpus. Il s'agit de noms qui présentent un caractère abstrait ou non-spécifique et fonctionnent en tant qu'agents de cohésion - grammaticale et lexicale – et leur référence est saturée par des éléments du texte en anaphore ou cataphore.

Notre travail s'inscrit dans la linguistique de corpus. Nous avons analysé quantitativement et qualitativement les emplois de deux noms généraux – *geste* et *démarche* – dans le corpus du *Monde* en utilisant un concordancier qui permet d'extraire le motif recherché entouré de ses contextes gauche et droit. Les emplois hétérogènes des noms généraux n'ont pas permis d'automatiser le processus d'annotation. Le travail a donc été réalisé manuellement.

L'analyse effectuée a permis d'étudier le processus référentiel et cohésif desdits noms. Nous avons proposé une typologie de leurs emplois cohésifs.

Malgré leur caractère abstrait, ces objets sans référence peuvent avoir une connotation subjective et faire l'objet d'une personnalisation. L'analyse a permis de mettre en évidence la nature « subjective » de certains des emplois de *geste* et de *démarche* dans le corpus du *Monde*.

Le travail sur les noms généraux se trouve dans (Adler & Eshkol-Taravella 2012, à paraître).

5.2. Noms généraux : définition

La catégorie des « noms généraux » a été proposée par (Halliday et Hasan 1976). Ces dernières années, la question des noms généraux a été reprise par la linguistique de corpus. Ainsi, (Schmid 2000)[55] et (Mahlberg 2005) ont testé les fréquences d'emploi des noms généraux dans des corpus écrits anglophones. En linguistique française, (Legallois 2008) a étudié quelques caractéristiques des « noms sous-spécifiés » en français. (Cappeau et Schnedecker 2014) ont étudié les noms généraux *gens, personne(s), individu(s)* dans le discours. Parmi les travaux en TAL, on peut mentionner (Kolhatkar *et al.* 2013a,b) qui se sont assigné pour objectif la résolution d'anaphores en anglais ou (Rose *et al.* 2014) qui se sont intéressés à la détection automatique de ce type de noms.

Les noms généraux sont des mots comme *chose*, *fait*, *idée*, *problème* ou *question*, qui fonctionnent en tant qu'agents de cohésion à la fois grammaticale et lexicale. Ils refèrent à d'autres éléments du texte (une proposition syntaxique, une phrase ou des unités discursives plus larges) en anaphore ou cataphore et leur atrribuent un label en caractérisant ainsi leur contenu. Ces noms portent le caractère abstrait ou non-spécifique, ce qui contraint à recourir au contexte – droite ou gauche – afin d'en décider l'imprégnation lexicale.

Cette partie concerne ce phénomène et s'inscrit dans le domaine de linguistique de corpus. Il constitue un travail exploratoire dans la perspective du traitement automatique des noms généraux en discours. Nous nous sommes intéressées avec Silvia Adler aux emplois cohésifs

[55] Signalons que Schmid parle de « shell nouns ».

de *démarche* et *geste* en discours médiatique. Les objectifs étaient de quantifier ces emplois dans la presse française, de typer les entités auxquelles réfèrent ces noms et d'esquisser une première analyse de leur contexte proche. Cette analyse a permis de constater que ces noms neutres sont souvent accompagnés d'expansions portant une émotivité ou une évaluation subjective.

5.3. Polysémie des noms *geste* et *démarche*

Geste et *démarche* s'inscrivent avant tout dans le domaine du « mouvement corporel » :

> *Soudain, tout s'est mis à tourner autour de moi. Les soldats, le thé, les panneaux, les gens, les mots. Il fallait descendre, vite. Le vélo démonté dans le coffre, j'ai sauté sur le siège passager du 4 × 4 et Yangjor a mis le contact. D'un* ***geste*** *tremblant, j'ai salué les militaires qui retournaient à leur poste, prêts au combat pour que les roses continuent de pousser de ce côté de la frontière.* (Le Monde 2013-08-26)

Lorsque le substantif masculin *geste* est accompagné d'un complément, il peut dénoter, selon le TLF, « le sentiment » ou « la réaction de l'auteur du geste ». Le mot connaît d'autres emplois dont un emploi figuré où il équivaut à « action ». Le dictionnaire Larousse en ligne ajoute au sens corporel celui d'« action remarquable qui frappe par sa générosité, sa noblesse etc. ».

Pour ce qui est du sens fondamental du nom *démarche*, le TLF cite « la façon de marcher », mais, au figuré, *démarche* signifie « manière d'avancer dans un raisonnement » ou « manière de penser ». Le dictionnaire Larousse en ligne propose également « manière d'agir » et « tentative faite auprès de qqn » comme dans l'exemple suivant :

> *En 2006, la droite au pouvoir avait complexifié par la loi le droit à vivre en famille de ces « conjoints de Français ». En plus du reste des* ***démarches*** *à effectuer pour venir en France, en 2006, une loi leur a imposé d'obtenir un « visa de long séjour ».* (Le Monde 2013-08-21)

Nous nous intéressons aux autres emplois de ces noms, aux emplois cohésifs, c'est-à-dire à un fonctionnement grammatical et lexical qui se saturent en référence par le truchement d'autres éléments du texte en anaphore ou cataphore souvent éloignés de la phrase où ils se trouvent employés.

5.4. Analyse quantitative de geste et démarche dans le corpus médiatique

(Francis 1994) classe le nom général *move* parmi les labels les plus populaires dans le corpus du Times. Reprenant ce constat, nous avons vérifié la fréquence d'emploi de ses correspondants *geste* et *démarche* dans la presse française. Le Tableau 17 donne un aperçu de leur fréquence d'utilisation dans un corpus médiatique du français.

Ce tableau n'est pas suffisant pour analyser objectivement la fréquence des noms généraux *geste* et *démarche* dans le corpus. Tout d'abord, la fréquence absolue affichée sur les sites consultés ne tient pas compte de la taille du corpus ce qui ne permet pas d'avoir une vue globale sur le phénomène. Ensuite, les noms étudiés ont des emplois multiples qui ne marquent pas tous la cohésion (voir la section 5.3).

Corpus	**Europresse**[56]		**Est Républicain**[57]
Rubrique			
Période	7 jours	30 jours	deux mois
fréquence *geste*	1374	6039	1068
fréquence *démarche*	2057	9779	1448

Tableau 17 : Fréquence absolue de *geste* et *démarche* dans la presse française (toutes rubriques confondues)

Pour étudier les emplois de deux noms dans la presse d'une manière plus objective, nous avons utilisé deux sous-corpus extraits du Monde durant une période indéterminée en 1988 et durant un mois du 19 août au 19 septembre en 2013. Les emplois cohésifs et non cohésifs ont été filtrés manuellement. Le Tableau 18 présente les fréquences absolues et relatives des ces mots dans les deux corpus. Les fréquences sont présentées selon leur nature cohésive ou non cohésive.

En premier lieu, on constate que les noms étudiés ont une fréquence relative quasi stable dans les deux corpus qui varie entre 0,006% et 0,008% par rapport à la taille du corpus. Le Tableau 18 montre aussi que le nom *démarche* est plus souvent employé en tant que nom cohésif dans la presse (75%-85% des emplois de ce mot sont cohésifs) par rapport à *geste.* Le mot *geste* présente moins d'homogénéité. On peut présupposer que *geste* est plus ancré sémantiquement, c'est-à-dire qu'il conserve plus que *démarche* son sens d'origine de mouvement corporel. Cette hypothèse se vérifie par la comparaison entre les deux rubriques du journal Le Monde *: Politique* et *Sport*. Il s'avère que la grande majorité des emplois repérés pour *geste* et *démarche* dans la rubrique *Politique* du journal Le Monde en ligne durant la tranche temporelle allant du 19 août au 19 septembre 2013 représente des emplois en tant que noms généraux assurant une cohésion textuelle. Dans la rubrique *Sport* on remarque une majorité d'emplois non cohésifs. *Geste* est utilisé dans les deux types de corpus plutôt en tant que nom non général ; en ce qui concerne *démarche,* sa distribution montre un emploi cohésif prédominant dans l'écrit médiatique.

5.5. Analyse des emplois des noms généraux

5.5.1. Proposition de typologie des emplois fondée sur l'analyse du contexte

L'étude du processus référentiel et cohésif desdits noms a été décrite dans (Adler et Eshkol 2013). Je me contenterai de présenter ici une synthèse de ce travail. En nous fondant sur l'analyse manuelle du sous-corpus extrait de la revue Le Monde 1998 à l'aide du concordancier Lextutor, nous avons proposé six schémas de cohésion observés.

- (DET) + (Modifieur) + NG = > action unique

  *Le procureur général d'Arménie Genrik Khatchatrian a été **assassiné**, jeudi 6 août, dans son bureau à Erevan par l'un de ses collègues qui s'est donné la mort juste après*

[56] http://www.europresse.com/
[57] http://www.cnrtl.fr/corpus/estrepublicain/

*avoir commis **son geste**, a-t-on annoncé de source officielle à Erevan. Aucune explication officielle n'a été donnée pour expliquer ce meurtre et ce suicide.* (Le Monde 1998 : 038)

Corpus	**Le Monde 1988**[58]	**Le Monde 2013**	
Rubrique	?[59]	*Politique*	*Sport*
Période	?[60]	19 août 2013 - 19 septembre 2013	
Taille	1110392 mots	277614 mots	105559 mots
Geste			
fréquence totale absolue	85	16	9
fréquence totale relative	0,007%	0,006%	0,008%
emplois cohésifs	29	12	1
emplois non cohésifs	56	4	8
Démarche			
fréquence totale absolue	98	16	1
fréquence totale relative	0,008%	0,006%	0,0009%
emplois cohésifs	87	12	1
emplois non cohésifs	11	4	0

Tableau 18 : Fréquence des emplois cohésifs vs non cohésifs

- (DET) + (Modifieur) + NG => action composite

  *Jim Barsdale entend ainsi « stimuler les énergies créatives de la communauté d'Internet et atteindre des niveaux d'innovation sans précédent sur le marché des navigateurs ». **L'entreprise ouvrira un site où les internautes pourront télécharger ce code source, communiquer leurs améliorations et débattre du sujet. Les développeurs seront libres de « modifier » et de « redistribuer » Communicator 5 0, qui, outre l'outil de navigation, intègre la gestion du courrier électronique, la participation aux forums, l'édition de pages sur la Toile...** Originale dans le monde de l'informatique commerciale, **cette démarche** permet à Netscape de redorer son image sans pertes considérables.* (Le Monde 1998 : 008).

- DET + NG pl. => plusieurs actions

  ***Mettre en prison 49 personnes** soupçonnées d'avoir participé au massacre, **changer de ministre de l'intérieur**, et **écarter le gouverneur de la province du Sud** ont constitué les premières réponses, suivies bientôt par **la mise en cause d'un haut policier** accusé par ses hommes d'avoir aidé à transporter les armes des assassins,*

[58] Ce corpus a été consulté sur le lien : http://www.lextutor.ca/concordancers/concord_f.html à l'aide du concordancier Lextutor.
[59] Les rubriques ne sont pas renseignées sur le site de consultation
[60] La période n'est pas renseignée sur le site de consultation.

*puis par **l'arrestation de 29 policiers** qui, le 12 janvier, à Ocosingo, au Chiapas, avaient tiré sur une manifestation d'Indiens, tuant une femme. **Autant de gestes** qui se veulent la démonstration que, cette fois, l'impunité n'est plus de mise. Mais, pour le répéter souvent, le président mexicain sait qu'il lui faut régler le problème au fond. Or, sur ce terrain, ses moyens sont singulièrement limités.* (Le Monde 1998 :007)

- (DET) + NG => action + état résultant ou moyens + fin

*Par ailleurs, certaines entreprises cherchent déjà à s'affranchir des contraintes de la future loi. Comment ? En s'apprêtant à refuser les aides de l'Etat, qualifiées d'« argent sale » par le directeur des ressources humaines d'une grande entreprise de la métallurgie. […] **ce refus de l'aide de l'Etat permettrait aux entreprises de réduire le temps de travail à 35 heures en retirant certaines pauses du calcul de l'horaire de travail effectif**, **démarche** à laquelle s'oppose Mme Aubry dans l'argumentaire transmis aux parlementaires socialistes.* (Le Monde 1998 : 025)

- (DET) + (Modifieur) + NG => action unique ou composite + contextualisation (« lexicalisation ») « appositive »

*Le gouvernement sri-lankais a annoncé, mercredi 28 janvier, que les célébrations du cinquantenaire de l'indépendance du pays, prévues début février, auraient lieu dans la capitale, Colombo, et non plus à Kandy où un attentat-suicide de la guérilla tamoule a fait 16 morts, dimanche 25 janvier. **Les Tigres de libération de l'Eelam tamoul avaient fait exploser un camion piégé devant le temple de la Dent à Kandy** (centre), le lieu le plus sacré des bouddhistes sri-lankais, dans **un geste spectaculaire de défi** aux autorités sri-lankaises en guerre contre la guérilla tamoule.* (Le Monde 1998 :046)

- (DET) + (Modification) + NG => action unique ou composite (en anaphore ET en cataphore)

*[…] le chef de l'Etat a félicité et remercié Jean Paul II pour sa visite et ses déclarations. Mais, pour couper court à toute spéculation sur une éventuelle ouverture du régime, il a indiqué que Cuba « croit en ses idées » et « défend de manière inamovible ses principes ». C'est seulement lors de son départ à l'aéroport, dimanche soir 25 janvier, que le pape **a exprimé sa condamnation du blocus économique auquel Cuba est soumis par les Etats-Unis**. Ce n'est pas la première fois que Jean Paul II fait état de **sa réprobation des sanctions « en général » et de l'embargo contre Cuba en particulier, GESTE** tant attendu par les autorités cubaines, mais, cette fois, **il l'a fait in situ, en des termes tranchants et définitifs, dénonçant** « la pauvreté matérielle et morale, dont les causes peuvent être les injustes inégalités, les limitations des libertés fondamentales, la dépersonnalisation, le découragement des individus et les mesures économiques restrictives imposées de l'extérieur du pays, injustes et éthiquement inacceptables. »* (Le Monde 1998 :033).

Notre étude a permis de mettre en valeur la richesse des processus cohésifs instaurés par les noms *geste* et *démarche*. Ainsi, ces noms généraux peuvent, lorsqu'ils sont au singulier, référer à un procès unique ou composite dont les différentes phases constitutives peuvent être ou non rattachées par un lien de causalité.

5.5.2. Geste et démarche : organisateurs discursifs à potentiel affectif

L'analyse des noms généraux *geste* et *démarche* a montré qu'ils peuvent être porteurs d'un discours subjectif. Dans ce cas ils sont modifiés ou expansés par des unités lexicales de nature

affective ou évaluative. Ainsi, dans les exemples suivants, le nom *geste* est modifié par les adjectifs *fort* ou *beau* apportant le jugement et l'appréciation sur le fait et l'événement annoncé :

> *Pour les entreprises, les exonérations liées aux heures supplémentaires ont constitué un véritable effet d'aubaine. Le nombre moyen d'heures supplémentaires trimestrielles, de 7 au troisième trimestre 2007, est monté jusqu'à 11,3, avant de redescendre à 9,7 au premier trimestre 2013. La suppression de ce dispositif avait été un des premiers* ***gestes forts*** *de la majorité élue en 2012, destiné à rompre avec un symbole du "travailler plus pour gagner plus".* (Le Monde 2013-09-03)

> *Les Girondins n'ont pu que réduire la marque sur une très jolie frappe du gauche d'Obraniak. Mais ce* ***beau geste*** *n'a pas calmé la colère froide de l'entraîneur bordelais Francis Gillot.* (Le Monde 2013-09-01)

Les journalistes, à travers ces adjectifs, semblent transmettre leur perception des faits et des événements. Les modifieurs peuvent avoir une connotation péjorative comme dans l'exemple suivant :

> *Parce qu'il est en responsabilité, François Hollande décrète la pause fiscale pour ne pas risquer de casser la fragile reprise de la croissance qu'il dit entrevoir depuis le 14 juillet et qui, seule, peut l'aider à réduire les déficits sans faire trop de casse sociale. Cette pause pourrait n'être qu'un* ***geste opportuniste****, en attendant des jours meilleurs.* (Le Monde 2013-09-04)

Cette évaluation de *geste* ou *démarche* peut être accompagnée de l'expression d'un sentiment et peut se manifester d'une manière moins directe sans le lexique approprié mais à travers un contexte élargi. L'exemple qui suit est extrait d'un entretien avec le président de l'UMP en 2013, Jean-François Copé, où celui-ci ne cesse de s'en prendre à la méthode du Président de la République dans la gestion de la crise syrienne. Le journaliste essaie à un moment de changer la tonalité ou le sujet en proposant à Copé de reconnaître que tout n'est pas à critiquer dans la gestion du gouvernement actuel et que la réforme des retraites est une *démarche habile :*

> *Depuis un an, vous accusez la gauche d'être dans le déni. Or, elle vient d'annoncer une réforme des retraites sans bloquer le pays. Reconnaissez-vous* ***une forme d'habileté*** *dans la* ***démarche*** *du président de la République ?*

la réponse est immédiate et l'interviewé traite la *démarche* en question de *lâcheté* :

> *Ce qui vient d'être annoncé n'est pas une réforme, c'est une augmentation de taxes. Une de plus ! Les jeunes Français, auxquels François Hollande avait dédié son élection, seront les premières victimes de cette* ***lâcheté****.* (Le Monde 2013-09-02)

5.6. Bilan

5.6.1. Compte rendu

Objectif visé	Etudier le processus référentiel et cohésif de deux noms généraux *geste* et *démarche*, tous deux désignant aussi « le mouvement corporel » dans le corpus médiatique du Monde
Méthodologie	Utilisation des outils de linguistique de corpus : concordanciers, tableur Excel pour pouvoir analyser le contexte des emplois cohésifs
Données traitées	Deux rubriques *Politique* et *Sport* du corpus du Monde extraites du Web du 19 août au 19 septembre 2013. Corpus Le Monde 1998 disponible en ligne avec le concordancier Lextutor
Difficultés rencontrées	Impossibilité d'avoir un accès libre au corpus Le Monde en ligne pour pouvoir le traiter Recours à un outil qui permette d'aspirer ce corpus depuis le Web afin de le traiter et de le soumettre au concordancier. Impossibilité d'automatiser le processus d'annotation à cause des emplois hétérogènes des noms généraux
Résultats	Analyse quantitative et qualitative des emplois cohésifs de deux noms *geste* et *démarche* dans le corpus du Monde Elaboration de la typologie du processus référentiel de ces noms : six schémas de cohésion Etude comparative quantitative des emplois cohésifs et non cohésifs de ces noms dans les deux rubriques *Politique* et *Sport*
Contraintes et limites	Typologie établie n'est pas implémentable actuellement Travail manuel
Originalité et apport du travail	Mise en valeur de la richesse des processus cohésifs instaurés par les noms *geste* et *démarche* : - possibilité de référer à un procès unique ou composite dont les différentes phases constitutives peuvent être ou non rattachées par un lien de causalité. - valeur subjective de certains emplois cohésifs lorsque ces noms neutres sont accompagnés d'expansions portant une émotivité ou une évaluation subjective. Comparaison entre les deux noms dans les deux rubriques : *geste* est plus ancré sémantiquement et conserve plus que *démarche* son sens d'origine de mouvement corporel.

5.6.2. Perspectives et travaux en cours

Ce travail se poursuit suivant plusieurs perspectives. En premier lieu, en linguistique de corpus, nous avons comparé les emplois cohésifs et non cohésifs de deux noms *geste* et *démarche* dans les huit rubriques du corpus Le Monde : *Education, Idées, International, Politique, Sciences, Société, Sport, Technologies*. Le corpus extrait grâce au concordancier a été filtré et les emplois non cohésifs de ces mots en ont été retirés. Le corpus restant a été annoté en emplois subjectifs. Les résultats d'une étude en cours permettront d'affiner les emplois cohésifs des noms *geste* et *démarche* et d'observer quantitativement et qualitativement les emplois dits subjectifs quand ces noms sont des organisateurs du discours avec un potentiel affectif. En deuxième lieu, je veux élargir cette étude aux autres noms généraux comme *approche, dispositif, initiative, méthode, stratégie etc.* Ce travail a été entamé dans le cadre du cours « Traitement de l'information » en L3 où les étudiants ont analysé les différents contextes d'emplois de ces mots à l'aide des outils comme AntConc et TXM. Enfin, j'aimerais me consacrer à l'exploitation de la typologie proposée pour la détection automatique de la cohésion dans le discours afin de contribuer à la résolution des problèmes de coréférence anaphorique. L'objectif est de permettre, d'une part, la détection automatique des segments auxquels réfèrent les noms généraux et, d'autre part, la distinction automatique entre leurs emplois cohésifs et non cohésifs. Pour cela, le corpus d'une taille plus importante sera annoté. Lorsqu'il s'agit d'un emploi cohésif, le segment auquel le nom réfère sera annoté également. Ce corpus annoté pourrait permettre d'identifier les constructions caractéristiques qui serviront à désambiguïser automatiquement l'emploi cohésif du nom général en utilisant entre autres la technique de l'apprentissage automatique supervisé. Des travaux de cet ordre ont déjà été réalisés pour l'anglais (Kolhatkar *et al.* 2013a,b) et pour le français avec les noms *problème* et *solution* (Rose *et al.* 2014). Notre démarche s'inscrit dans ce cadre mais les mots traités sont plus polysémiques et donc plus difficiles à traiter en TAL.

6. Conclusion : perspectives, réflexions et travaux futurs

Le recours croissant de la linguistique informatique aux méthodes probabilistes pose le problème de la légitimité de la présence du linguiste dans ce domaine et plus précisément du rôle qu'il peut jouer aujourd'hui dans le TAL. Ce mémoire est la réponse donnée à cette question par l'explicitation de mes propres recherches, en montrant comment compétences linguistiques et informatiques peuvent contribuer aux recherches dans ce domaine.

Je commencerai cette partie en décrivant la méthodologie appliquée dans mes recherches. Je présenterai ensuite une synthèse de mes travaux. Je développerai à la suite quelques perspectives de recherches actuelles et futures et je terminerai par certains points liés à mon activité d'enseignant.

6.1. Méthodologie de la recherche

Internet et les nouvelles technologies ont relancé l'intérêt pour l'étude et l'analyse des textes dialogiques écrits (blog, forum, chat, réseaux sociaux etc.) et oraux. Ces formes de communication caractérisées entre autres par l'expression des sentiments et des opinions se sont imposées via le Web et l'innovation sociale qu'elles représentent dans les nouvelles possibilités offertes par la technologie exigent un traitement spécifique. Le TAL a un rôle important à jouer dans l'accès, l'exploitation et l'analyse des données.

La première étape vers le développement de systèmes capables de repérer et analyser l'information consiste à créer des ressources annotées. C'est un thème central de mes travaux. Ma démarche est résolument empirique, guidée par les observables issus des corpus. Elle est fondée sur la préservation des spécificités linguistiques des corpus traités et sur la prise en compte de la variation linguistique qui y est présente. Cette méthodologie suivie dans tous mes travaux, quel qu'en soit le thème ou l'objet, est décrite dans la Figure 7. Elle procède en quatre phases :

- Observation et analyse manuelle du corpus : l'objectif est d'acquérir une familiarisation avec les données traitées et de recenser des variations dans les occurrences. De ce repérage découle une réflexion programmatique sur l'étendue des phénomènes à annoter, leur classification et l'identification de leurs marqueurs formels ;

- Modélisation : l'analyse préalable des données permet de modéliser l'information à étudier. L'opération consiste à établir une typologie sous la forme d'un jeu d'étiquettes correspondant à la nature du corpus, c'est-à-dire en tenant compte de ses spécificités d'une part et des résultats à atteindre d'autre part.

- Etablissement de la technologie adaptée : au moment de la prise de décision concernant la technologie, le choix est déterminé par le respect des données linguistiques et des contraintes que ces données imposent au traitement automatique ainsi que par la finalité déclarée. Il s'agit généralement de trancher entre deux options, soit le développement d'un nouvel outil, soit l'adaptation d'un outil existant. C'est à ce moment que se décide la méthode de l'annotation automatique : symbolique fondée sur des règles et/ou statistique exploitant la technique de l'apprentissage automatique ;

- Implémentation du schéma d'annotation défini afin de procéder à une analyse quantitative et qualitative des résultats.

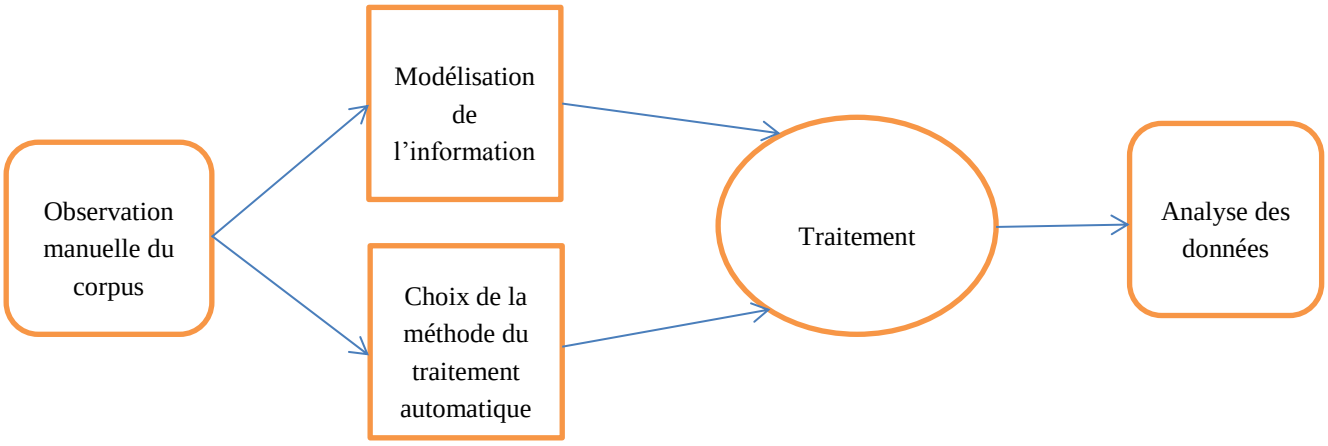

Figure 7 : Méthodologie

6.2. Synthèse

Les travaux que je mène depuis une douzaine d'années concernent principalement le développement et l'exploitation des corpus. Je considère que le travail préalable au traitement et à l'exploitation du corpus constitue, au même titre que les résultats quantitatifs, un apport appréciable en TAL. La réflexion sur la nature des données traitées semble parfois sous-estimée. Constitution des données et traitement sont indissociables et doivent être suivis de la collecte jusqu'à la diffusion du corpus. Les outils ne sont pas préexistants aux données. Ils doivent répondre aux besoins particuliers liés aux corpus à traiter et/ou analyser. Ce principe du respect des données traitées apparaît dans tous mes travaux décrits dans ce mémoire et synthétisés dans le schéma de la Figure 8 qui sera développé dans la partie qui suit.

Annotation

Mes travaux portent sur l'annotation[61] des corpus. Je considère ce processus comme une sorte de modélisation de l'information. On passe du niveau du langage naturel vers un formalisme capable d'être traité informatiquement. L'ajout des étiquettes enrichit le corpus et le rend traitable par la machine car l'information annotée peut faciliter sa consultation et son analyse automatiques.

Le principe que je suis est le respect de la nature des données. Mes compétences et intuitions linguistiques enrichissent beaucoup le travail informatique. De ce fait, je valorise l'apport de la linguistique qui n'est pas toujours suffisamment exploité dans les recherches en TAL.

Du corpus normalisé écrit vers les corpus hors normes

Les corpus que j'ai pu traiter et que j'ai décrits dans ce mémoire sont au nombre de trois : Le corpus oral ESLO, le corpus Web des titres de cartes géographiques et le corpus médiatique du *Monde*. La flèche verticale de la Figure 8 montre la différence de nature de ces corpus et la

[61] Je vais souligner les mots clefs faisant partie du schéma de la Figure 8 : Schéma synthétique de tous mes travaux

progression du traitement du corpus normalisé écrit du Monde vers des corpus moins normalisés et hors norme. Ces corpus sont difficiles à traiter de façon automatique. Il s'agit tout d'abord du corpus du Web, une sorte de combinaison entre discours oral et écrit. Un autre corpus non standard est le corpus oral transcrit ESLO où l'absence de signes typographiques, la présence de disfluences interrompant le flux de la parole, les chevauchements entre locuteurs etc. rendent difficile un traitement automatique dès lors que les outils à disposition ont le plus souvent été développés pour des corpus écrits normalisés. Une réflexion approfondie sur les outils à développer ou à adapter, sur les jeux d'étiquettes à établir etc. est indispensable pour éviter certains biais de l'analyse.

Le principe sur lequel se règlent mes travaux consiste à mettre à profit la nature du corpus et le surcroît d'informations linguistiques qu'il livre afin d'améliorer les techniques de traitement automatique. L'apport général de mes travaux sur les corpus hors normes est le développement d'outils permettant leur annotation.

Cette orientation a été mise en œuvre dans mes travaux portant sur l'annotation syntaxique du corpus oral, sur le repérage et l'annotation des entités nommées et sur le discours oral et ses phénomènes idiosyncrasiques avec pour finalité le traitement automatique.

6.2.1. **Annotation syntaxique**

Deux types d'annotation ont été effectués : l'annotation syntaxique (étiquetage morpho-syntaxique et chunking) et l'annotation sémantique.

L'annotation syntaxique d'un corpus oral consiste dans l'affectation d'une information sur la catégorie syntaxique, le genre, le nombre etc. des mots identifiés et dans l'étiquetage des syntagmes minimaux (chunks). Les unités de traitement sont les mots isolés dans le premier cas et les syntagmes dans le deuxième. La transcription d'un corpus oral s'effectuant sans signes typographiques, avec des disfluences qui interfèrent dans le flux discursif, des chevauchements entre les tours de parole etc., l'automatisation des procédures s'avère bien plus complexe qu'à l'écrit.

L'originalité de ce travail tient à l'application des techniques d'apprentissage automatique à des corpus oraux et à la définition d'un jeu d'étiquettes qui permette d'intégrer les phénomènes caractéristiques de l'oral.

6.2.2. Annotation sémantique

J'entends par l'annotation sémantique le repérage, l'extraction et l'analyse de l'information annotée de nature sémantique et/ou pragmatique.

L'annotation sémantique porte sur les trois corpus étudiés. Elle est confrontée à la notion de subjectivité (voir Figure 8

6.2.2.1. Entités nommées

L'annotation des entités nommées (groupes nominaux renvoyant vers les personnes, lieux, organisations etc. et expressions numériques) est, pour moi, le premier cas de figure de l'annotation sémantique.

Mes travaux se caractérisent par le rapprochement de la notion d'entité nommée avec celle de subjectivité.

A première vue, une notion comme celle d'entité nommée semble être très objective car elle renvoie à un référent unique. Selon (Ehrman 2008), « on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus ». Pourtant,

la désignation d'une personne ou d'un lieu est un processus social réapproprié subjectivement, la signification d'un anthroponyme ou d'un toponyme, c'est-à-dire la référence à une personne ou à un lieu donnés, est déterminée par le système personnel et donc subjectif du locuteur. J'ai ainsi confronté l'usage des entités nommées (noms de lieux, de personnes, d'organisations etc.) avec une appréciation de la subjectivité latente dans leurs emplois.

Les travaux sur l'anonymisation du corpus oral ESLO et sur le repérage de l'information spatiale dans le corpus Web des titres de cartes m'ont permis d'observer ce passage de l'entité nommée vers l'entité à connotation subjective. Mes recherches ont montré que l'emploi des entités nommées dans un discours de nature subjective comme c'est le cas des entretiens oraux ou du corpus Web de titres de cartes proposés par les internautes, s'accommodent d'une interprétation subjective.

6.2.2.1.1. Faisceau d'indices d'identification

Dans ce cadre, j'ai participé à l'essai d'anonymisation automatique des transcriptions du corpus oral ESLO afin d'en assurer la mise à disposition sur le Web. J'ai défini une méthode de repérage automatique de l'information personnelle qui, déclarative ou par recoupement, permet d'identifier un locuteur, ce que j'ai appelé le *faisceau d'indices*.

La reconnaissance des entités nommées ne suffit pas à repérer toutes les informations concernant la personne. Les entités nommées présentes dans le discours doivent avoir un lien avec le locuteur ou la personne qu'il mentionne dans son discours pour devenir un indice d'identification. Aux entités nommées identifiantes s'ajoutent d'autres éléments assez hétérogènes désignant les différentes informations personnelles sur la personne : événements, activités sociales, loisirs, maladies, handicap etc. qui peuvent au même titre que le travail, la famille donner les informations sur le locuteur ou la personne dont on parle.

Je considère que lorsqu'un élément dans le discours permet d'identifier le locuteur, il n'est plus neutre car il est perçu par l'interlocuteur/auditeur d'une certaine manière. La notion de subjectivité est intrinsèquement liée ici, de mon point de vue, avec celle de la perception.

Les apports du travail sont :

- la nature des éléments annotés, qui n'ont jamais, à ma connaissance, fait l'objet d'une étude particulière dans la littérature scientifique

- une approche fine de l'anonymisation, contrairement à nombre de travaux du TAL, qui ne se limite pas au repérage automatique des entités nommées

- la démonstration des limites du traitement automatique de ce processus

- la constitution d'un corpus oral annoté en entités nommées et indices d'identification

6.2.2.1.2. Lieux (en général) et lieux subjectifs

L'annotation des entités nommées est également au centre du travail effectué sur le corpus écrit issu du Web des titres de cartes. Il s'agit de repérer les noms de lieux dans ce corpus. Le corpus traité composé des titres de cartes créés par les internautes est très particulier. Les noms y sont désignés d'une manière très variée et parfois personnelle et intime. Les entités nommées « classiques » ne peuvent pas couvrir une telle diversité de désignations de lieux. Il faut ajouter à des entités nommées connues les nouvelles désignations de lieux lorsqu'ils sont

personnalisés, appropriés, inventés etc. par le locuteur. Nous avons nommé ces lieux *lieux subjectifs.*

Malgré l'intérêt actuel croissant pour l'expression de la subjectivité dans le discours, les noms de lieu dits subjectifs n'ont jamais, à ma connaissance, fait l'objet d'une étude particulière.

L'apport de ce travail est la mise en évidence de cette classe de lieux, leur annotation et leur analyse.

6.2.2.2. Reformulation paraphrastique

La reformulation paraphrastique est un autre phénomène faisant l'objet de l'annotation sémantique. Le fait de reformuler son discours est un trait inhérent de l'expression orale dont les manifestations sont très différentes de celles qu'on rencontre à l'écrit. La reformulation paraphrastique peut être produite au niveau de plusieurs énoncés, mais, dans mon travail, je me suis contentée d'un énoncé (équivalent à un tour de parole). La taille de l'unité traitée a agrandi : des syntagmes dans le cas des entités nommées, je suis passée aux tours de parole.

Pour étudier ce phénomène et le repérer de manière automatique, j'ai proposé la constitution d'un jeu d'étiquettes multidimensionnel tenant compte de différents aspects linguistiques (modifications morphologiques, lexicales, sémantiques etc.). Les tours de parole contenant la reformulation ont été annotés par ce moyen ce qui a permis de contribuer au développement des règles de reconnaissance automatique de ces tours de parole.

Les apports de ce travail sont :

- le recours à une approche syntagmatique contrairement aux approches utilisées souvent en TAL pour effectuer cette tâche

- l'annotation multidimensionnelle

- le travail à l'intérieur d'un grand corpus oral quand beaucoup de travaux en TAL se limitent au repérage automatique des paraphrases dans des corpus homologiques, bilingues ou monolingues écrits

- l'extension de la notion de paraphrase qui ne se limite pas à l'équivalence syntaxique : les explications, précisions, définitions, résultats etc. sont aussi pris en compte

- la production d'un corpus oral annoté en reformulation paraphrastique

6.2.2.3. Discours « neutre »

Le troisième cas de l'annotation sémantique est représenté par mon travail sur les discours « neutres ». J'entends par discours « neutre », le discours standard ordinaire qui ne doit normalement pas avoir de connotation subjective. L'exemple de ce type de discours dans mes recherches est la recette d'un plat aussi banal que l'omelette et des noms généraux. Dans les deux cas, l'objet d'étude est tellement commun ou tellement abstrait qu'on peut supposer qu'il n'a plus de référent. On est en quelque sorte à l'opposé des entités nommées traitées ci-dessus. Malgré ces hypothèses, il s'est avéré que même ces objets sans référence peuvent avoir une connotation subjective et être personnalisés dans le discours.

L'analyse des discours « neutres » fait partie de mes travaux dans le domaine de la linguistique de corpus. Les outils du TAL ne sont presque pas utilisés. C'est un travail peu automatisé car il dépasse le niveau d'un énoncé. L'objet d'étude est le discours.

6.2.2.3.1. Recettes d'omelette

En comparant des recettes d'omelettes consignées par écrit avec celles que livre à l'oral, au cours d'une enquête, une centaine de témoins, j'ai typé, annoté et extrait les informations dans une perspective pragmatique (au cours de l'interaction) et cognitive (planification de la tâche, script…). Une fois résolue la question du contenu informatif, d'autres types d'informations appartenant à la communication orale comme les marques de l'énonciation, les éléments phatiques, la diversité des niveaux du discours (interpellations de l'auditeur, anecdotes etc.), les connecteurs ou encore les disfluences ont été étudiées en lien avec les différentes phases de l'explication.

Les recettes de cuisine écrites sont souvent utilisées pour qualifier les textes procéduraux. Ce travail sur l'oral a permis d'observer et de dégager les différences entre les deux modes de restitution. Dans le cas de la recette écrite, l'auteur suit un format prédéfini de la recette. A l'oral, on assiste à l'élaboration de la recette, faisant intervenir des permutations dans les opérations ou des oublis, des commentaires liées ou non à la tâche demandée, l'expression de sentiments et d'opinions, des refus de réponse, des implications de l'enquêteur etc. La recette, qui est « neutre » dans son principe, est reformulée en discours subjectif.

6.2.2.3.2. Noms généraux

L'objet de la dernière étude présentée dans ce mémoire concerne les noms généraux. On entend par « noms généraux » des noms abstraits comme *geste, démarche, tentative, objet etc.* qui renvoient à d'autres éléments dont la désignation est différente dans le texte. L'objectif a été d'analyser les emplois de deux noms généraux *geste* et *démarche* dans le corpus du Monde. Même s'il s'agit d'un corpus normalisé écrit, l'objet d'étude est complexe.

Le processus cohésif de ces deux noms a été analysé et six schémas de cohésion ont été proposés. L'analyse quantitative de ces mots dans les deux rubriques du corpus *Politique* et *Sport* a montré que *geste* est plus ancré sémantiquement, c'est-à-dire qu'il conserve plus que *démarche* le sens d'origine qui est le sien de mouvement corporel. Le dernier constat de cette étude concerne la subjectivité desdits noms. Le travail a montré qu'ils peuvent être porteurs d'un discours subjectif lorsqu'ils sont modifiés ou expansés par des unités lexicales de nature affective ou évaluative.

L'étude du processus cohésif de ces noms généraux peut être utile dans le domaine du TAL pour la résolution des anaphores. Le rapprochement fait entre les noms généraux et le discours subjectif n'a jamais été mentionné dans les recherches sur ce type d'unités et mérite une étude plus approfondie.

En conclusion,

Mes travaux contribuent à l'amélioration du traitement automatique des corpus oraux et au repérage et à l'analyse du discours subjectif. Mon regard en tant que linguiste a permis d'enrichir ces travaux par une annotation plus fine et de mettre en évidence certains éléments méconnus ou ignorés dans les recherches actuelles en TAL et/ou en linguistique.

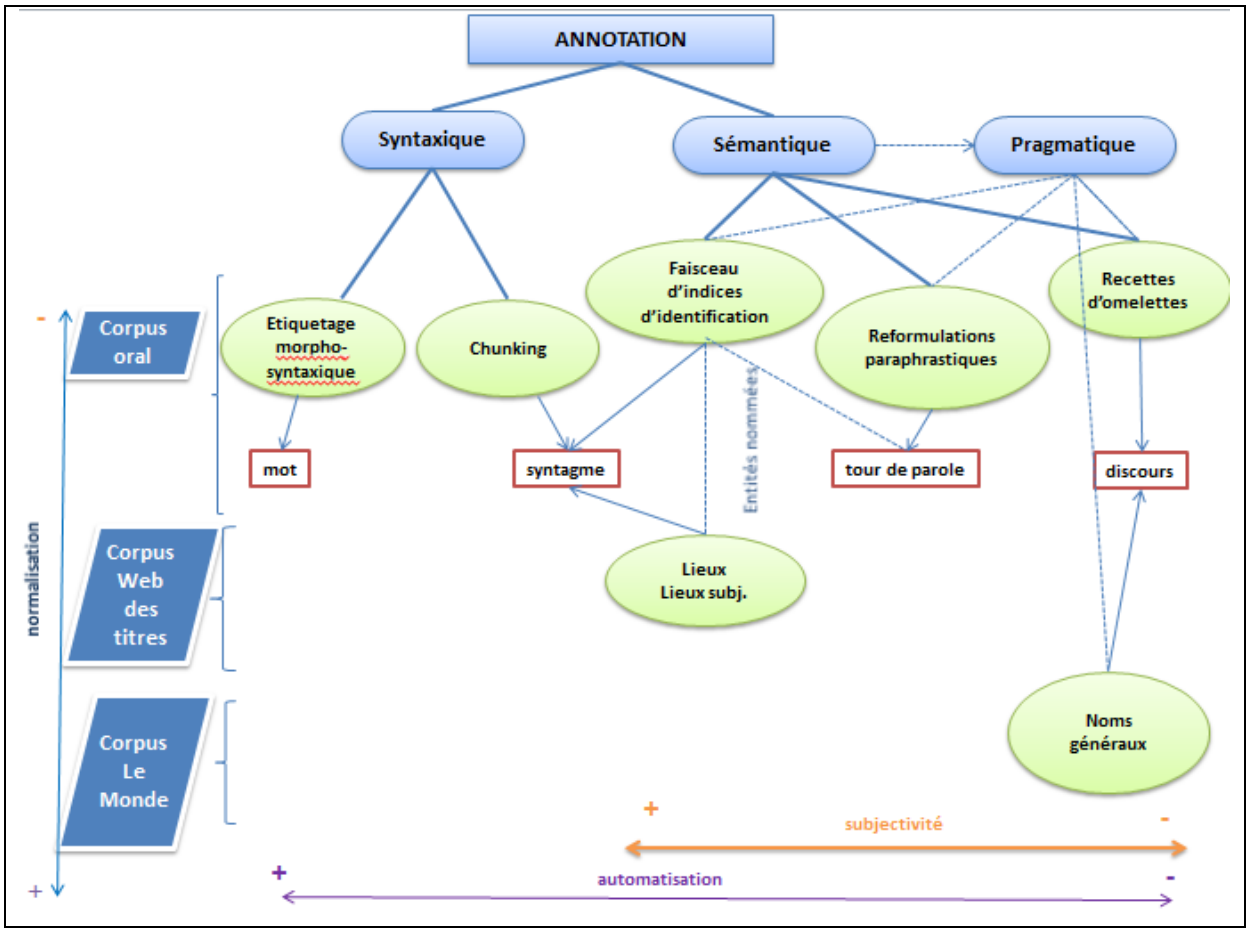

Figure 8 : Schéma synthétique de tous mes travaux

6.3. Perspectives

Mes réflexions sur la subjectivité ou sur la différence entre les discours oral et écrit se poursuivent actuellement.

6.3.1. Thématique de recherches : la subjectivité

C'est en travaillant sur l'anonymisation d'un corpus oral par repérage des informations identifiant le locuteur que j'ai abordé les questions sur la subjectivité dans le langage et sur son repérage automatique dans le discours. L'étude de la subjectivité et son expression dans le discours a été poursuivie dans plusieurs de mes recherches : subjectivité à partir des informations personnelles livrées par le locuteur, subjectivité dans la perception et l'appropriation des lieux, subjectivité dans les recettes de cuisine et enfin subjectivité exprimée à travers les noms généraux. Les travaux en TAL appelés *fouille d'opinions* (*opinion mining*) ou *analyse de sentiments* (*sentiment analysis*) s'intéressent aujourd'hui à différents aspects liés à cette notion : analyse et traitement du langage subjectif, construction et acquisition de ressources linguistiques, annotation et extraction de l'information subjective, modélisation du discours subjectif[62]. Dans une fouille de données et plus précisement dans la tâche visant le classement automatique des textes par l'apprentissage supervisé, les chercheurs se limitent généralement à une évaluation à trois degrés : positif, négatif, neutre. Ce fait a été constaté également par (Jackiewicz 2014) : « De nombreuses études en TAL concernées par la fouille d'opinions ou l'analyse de sentiments prennent comme point de départ la constitution de dictionnaires de mots « subjectifs », idéalement étiquetés en « positifs » et « négatifs ». Nombre de travaux en linguistique s'intéressent à des ensembles de lexiques dits psychologiques ou affectifs ou encore présentés comme renvoyant à des états privés ou à des attitudes propositionnelles. » L'auteure poursuit en expliquant que le langage est beaucoup plus subtil et nuancé. « Il n'existe pas de mécanisme grammatical spécifique ou de catégorie lexicale dédiée à l'évaluation. On atteste, en revanche, d'un grand nombre de procédés et de formes, réparties sur tous les plans de la structure linguistique, participant, dans une diversité d'acceptions, à des stratégies et des finalités discursives fort nombreuses.[…] Les frontières entre ce qui serait de l'ordre de l'appréciatif, de l'évaluatif, de l'axiologique ou de l'affectif sont mouvantes et incertaines. Car il s'agit de phénomènes graduels, parfois intriqués, sensibles au cotexte et à l'extra-linguistique, souvent régis par des systèmes complexes de normes relatifs aux locuteurs évaluateurs et aux cibles évaluées.[…] S'il est indispensable de prendre en compte le contexte, il ne suffit pas de se contenter de travailler sur des éléments contextuels artificiellement limités (fenêtres de mots, phrases...). La visée pragmatique et la situation de la production du discours jouent un rôle déterminant. » Mes travaux s'inscrivent dans cette optique. Il convient, selon moi, d'enrichir ces travaux par des approches plus linguistiques qui prennent en compte la nature du corpus et le contexte dans son sens le plus large.

Le traitement des indices de subjectivité reste au centre de mes recherches actuelles, que ce soit pour les entités nommées ou le repérage automatique des désignations de lieux sur le Web. L'annotation d'abord manuelle des lieux dit subjectifs permettra de constituer un corpus de référence pour leur apprentissage automatique et l'évaluation du module.

Le repérage de l'information spatiale dans le corpus des titres sera testé sur le corpus oral ESLO et sur le corpus des migrants « Le ventre de Marseille » (que nous sommes en train de constituer) afin d'y repérer les désignations de lieux et les sentiments auxquels ces lieux sont

[62] Ces thématiques ont été présentées dans la journée ATALA « Fouille d'opinions et analyse de sentiments » qui a eu lieu le 21 mars 2015 à Paris.

associés chez les habitants d'Orléans et de Marseille, mettant en évidence la dimension symbolique des désignations topographiques.

Je veux poursuivre mes recherches sur le repérage des informations personnelles dans le corpus oral. En premier lieu, on pourra exploiter les propositions de (Schnedecker 2015) sur les « noms d'humains » pour améliorer le repérage des entités nommées et les autres mentions anthropologiques en discours. Les propriétés décrites par l'auteure comme, par exemple, la variation de genre (*le/la concierge*) ou la pronominalisation par *lui*, pourraient être exploitées par un module de repérage automatique. Pour aider l'anonymisation manuelle, la distinction pourrait se faire entre les éléments les plus sensibles à l'anonymisation, c'est-à-dire ceux qui apportent une information plus importante et plus spécifique, et ceux qui sont plus généraux. Ainsi, pour distinguer entre les noms de famille rares comme *Eshkol* ou *Kanaan* et très répandues *Dupond* ou *Durand,* on pourrait s'appuyer, dans le cas du corpus des ESLO, sur une information concernant la fréquence d'un nom propre, éventuellement pondérée par des critères géographiques. De la même manière, le locuteur peut désigner son métier par un seul mot *enseignant* ou en précisant *professeur de physique.* Ce passage d'un seul nom à un groupe nominal plus étendu grâce aux modifieurs concerne n'importe quelle caractéristique (maladie, loisir etc.). On pourrait ainsi attribuer plus de poids à ces éléments sensibles à l'anonymisation ce qui diminuerait le nombre des indices candidats à l'anonymisation et de cette manière aiderait la validation manuelle. Pour un travail futur, j'envisage aussi d'étudier avec précision dans le corpus ESLO, tous les éléments anonymisés par une procédure manuelle afin d'affiner la typologie du faisceau d'indices.

Dans mes recherches sur le discours subjectif dans les médias, l'étude des noms généraux dans le corpus du *Monde* étendu est en cours de finalisation. Il s'agit d'une comparaison des emplois lexicaux dans les différentes rubriques : *Education, Idées, International, Politique, Sciences, Société, Sport, Technologies.* L'analyse de ce corpus annoté des emplois dits subjectifs des noms généraux *geste* et *démarche* permettra de procéder à leur examen quantitatif et qualitatif. Je veux aussi élargir cette étude aux autres noms généraux comme *approche, dispositif, initiative, méthode, stratégie etc.*

La problématique liée à l'expression de la subjectivité et à sa perception se retrouve dans la thèse de doctorat de Sandra Cestic que je co-encadre et qui, dans le cadre d'un contrat CIFRE avec l'entreprise ACATUS Informatique, modélise l'information sémantique afin de parvenir à une extraction automatique. L'objectif est d'analyser et de représenter les connaissances des salariés sur la perception de l'environnement de travail. A partir du corpus constitué par des entretiens avec les salariés, une étude des opinions, des avis, des appréciations et des évaluations en langage naturel spontané sera effectuée. Les résultats, qui aboutissent à une modélisation des données subjectives, sont rapportés aux données objectives obtenues par des mesures physiques.

Je co-encadrerai (puis co-dirigerai) une autre thèse sur contrat doctoral d'établissement. Il s'agit de la thèse d'Hélène Flamein intitulée « Entités nommées, opinions et sentiments analysés en corpus oral : repérage et analyse ». Ce doctorat se fonde sur les données d'ESLO. L'objectif est d'analyser la perception que les Orléanais ont de leur ville et de leur quotidien. Les entités nommées seront repérées dans les transcriptions des entretiens et leurs emplois seront analysés en contexte.

Je suis également impliquée dans le projet MODAL (MODèles d'Annotation de la modalité à l'oraL) qui vient d'être accepté par la MSH Val-de-Loire, et qui a pour objet l'annotation de l'attitude épistémique du locuteur vis-à-vis des représentations encodées dans l'oral dialogique.

6.3.2. Domaine de recherche : le discours oral / écrit

Le corpus oral reste central dans mes recherches qui s'inscrivent dans celles conduites au LLL. Le travail sur le français parlé et son annotation m'a conduit à une réflexion d'ensemble sur la modélisation des caractéristiques propres à l'oral : les disfluences (hésitations, faux départs, amorces, marqueurs discursifs etc.) où les reformulations font l'objet d'une étude particulière, les présentateurs, la segmentation, les commentaires personnels etc. Les résultats de cette modélisation peuvent être utilisés pour améliorer l'annotation syntaxique de l'oral par apprentissage automatique ou pour des applications plus industrielles comme le dialogue homme-machine, par exemple.

Les études sur l'oral se développent à travers différents axes correspondant à des questions nouvelles dans le champ de la recherche actuelle.

Pour améliorer les résultats de l'annotation syntaxique des corpus oraux, une réflexion est engagée en collaboration avec Isabelle Tellier sur le traitement commun d'un corpus oral et d'un corpus écrit. Si l'on veut développer un étiqueteur propre à l'oral en utilisant la technique de l'apprentissage automatique, il est nécessaire de disposer d'un corpus de référence de grande taille. C'est un travail de longue haleine qui nécessite la correction manuelle des étiquettes. Pour épargner du temps et disposer au final d'un corpus de référence d'une taille importante, nous voulons tester une méthode consistant à rapprocher de la transcription orale un corpus écrit. Ainsi, en ajoutant au corpus FTB disponible les caractéristiques du discours oral, on obtiendra un corpus de référence de très grande taille ce qui permettra un apprentissage plus efficace de l'annotation.

Pour effectuer cette transformation, une réflexion est menée sur les types de modifications à opérer dans le corpus écrit : la suppression des signes typographiques qui marquent la segmentation de l'écrit, l'ajout des présentateurs comme *c'est, il y a, on a etc.*, l'intégration des disfluences (répéter les déterminants, insérer les *euh* d'hésitation et d'autres interjections et des marqueurs discursifs, prévoir les amorces etc.). Les modifications peuvent porter aussi sur les différences stylistiques : les parenthèses de l'écrit pourraient être remplacées par les reformulateurs paraphrastiques comme *disons, c'est-à-dire, j'allais dire, par exemple etc.* Le contraste dans les tournures syntaxiques à l'oral et à l'écrit peut concerner les propositions relatives beaucoup moins présentes à l'oral. Le corpus ESLO étudié est un extrait des entretiens en face-à-face qui portent sur l'identité du locuteur, la vie à Orléans etc. Des variations dans le lexique sont à prévoir. Ainsi, les termes *marché commun, évolution économique* etc. présents dans les textes journalistiques traitant de politique ou d'économie peuvent être remplacés par des syntagmes plus fréquents dans la situation d'enregistrement. Les transformations doivent enfin concerner les symboles mathématiques comme %, =, + etc. qui sont écrits littéralement dans les transcriptions de l'oral *pourcent, égal à, plus* etc. Cette remarque concerne aussi les chiffres.

La réflexion sur la comparaison entre les discours écrit et oral transcrit sera poursuivie dans mes recherches sur les reformulations paraphrastiques. Ce processus consistant à éclaircir et faciliter la transmission et la compréhension de l'information diffère dans les deux cas du pont de vue du processus cognitif (Hagège 1985, Blanche-Benveniste *et al.* 1990). Pour analyser et comparer l'emploi de la reformulation paraphrastique à l'oral et à l'écrit, les trois corpus : ESLO (ESLO1 et ESLO2), le corpus médiatique du Monde et le corpus du forum de discussion Doctissimo, ont été annotés en reformulations. Ces trois corpus annotés permettent de faire des analyses comparatives sur la différence d'utilisation des reformulations dans des discours de nature différente. Je me propose de tester en collaboration avec Natalia Grabar l'efficacité des modèles développés pour la détection automatique des tours de parole avec la

reformulation paraphrastique dans le corpus oral sur d'autres types de corpus (journalistique et web). Nous ajoutons aux trois marqueurs étudiés *c'est-à-dire, je veux dire, disons* d'autres marqueurs : *notamment, autrement dit, en d'autres termes, en d'autres mots*, ce qui permettra d'avoir trois corpus riches en reformulations.

Une des limites inhérentes à mes études sur le corpus oral ESLO tient à la non prise en compte des informations provenant directement des fichiers sonores. Toutes mes recherches ont été effectuées à partir des transcriptions.

J'ai pour projet d'intégrer aux analyses les éléments prosodiques et acoustiques associés aux tours de parole, en commençant par le traitement des reformulations paraphrastiques dans le corpus ESLO. L'hypothèse de cette investigation est la détermination par ce moyen de la nature de différence entre les tours de parole avec reformulation paraphrastique et ceux qui, en ayant recours aux mêmes marqueurs, n'en présentent pas. Le filtrage introduit un critère innovant qui soulève des questions cruciales pour le TAL sur le plan de la multimodalité (écrit/oral).

Cette investigation est amorcée dans différents projets de recherche, auxquels je participe, portant sur le discours oral et son annotation. Il s'agit des programmes TEMPORAL (Annotation en référence et coréférence temporelle des corpus oraux) et DIASEMIE (Discrimination Automatique des Sens d'Emplois des Mots par l'Intonation). Après le projet MODAL (MODèles d'Annotation de la modalité à l'oraL), SEMORAL (Discrimination prosodique automatisée et sémantique de l'oral) vient de passer la première étape de pré-sélection d'ANR. Enfin, je fais partie de l'équipe du projet SegCor (Segmentation of oral corpora) soumis à l'appel à projets franco-allemand en sciences humaines et sociales par ANR.

6.4. Principes pédagogiques

Depuis 11 ans, je suis responsable du parcours TAL au département Sciences du langage de l'Université d'Orléans. Cette période a été marquée par quelques changements que j'ai pu observer et qui ont influencé mon enseignement.

Tout d'abord, j'observe le changement d'attitude de mes étudiants envers les nouvelles technologies. En 2003, lorsque j'ai commencé à enseigner, les étudiants appréhendaient l'ordinateur. Cette peur n'existe plus. Les doutes demeurent pour tout ce qui concerne la programmation. Un autre changement concerne cette fois le domaine des recherches en TAL où l'on note un recours croissant aux méthodes probabilistes même s'il semble que les entreprises privilégient encore les méthodes symboliques, ce qui peut encore évoluer. Ces deux tendances m'ont conduite à poser la question sur les compétences nécessaires en TAL qui doivent être transmises aux étudiants linguistes qui se spécialisent dans ce domaine. Je vais essayer de répondre à cette question à partir de mon expérience en tant que chercheuse et enseignante. Mes discussions avec mes anciens étudiants présents en entreprise me guident également dans le choix des contenus à délivrer aux étudiants.

Je distingue dans mon enseignement deux types de connaissances : passives et actives.

J'entends par connaissance passive une culture générale qui concerne les outils et les ressources libres disponibles. Il est nécessaire, me semble-t-il, pour un taliste de connaître des ressources disponibles : outils comme TreeTagger, CasEn, Melt etc. ; des dictionnaires ou lexiques (Lefff, Delaf, Delac, WordNet etc.), des corpus en ligne pour pouvoir les consulter en cas de besoin et les intégrer dans les nouveaux outils. Plusieurs sites recensent ces ressources, j'en citerai ici quatre :

- http://www.cnrtl.fr/
- http://deschamp.free.fr/exinria/divers/metalexis.html
- http://www.clt.gu.se/wiki/interactive-online-demos
- http://www.atala.org/-Outils-pour-le-TAL-

La base Frantext est fort intéressante pour montrer aux étudiants ce qu'est un corpus et ce qu'on peut en faire. La création de grammaires sous Frantext est un exercice très utile pour les étudiants afin de les initier aux bases des méthodes symboliques. Les compétences passives sont acquises au niveau de la Licence.

Les connaissances actives consistent dans le développement et l'utilisation des outils et demandent des compétences en informatique. C'est la raison pour laquelle elles sont données au niveau du Master. Elles peuvent être réparties en trois volets :

- constitution des corpus
- traitement des corpus
- analyse des données avec les outils (existants ou à développer)

La première étape liée à la constitution des corpus est primordiale car c'est à ce niveau que doivent être posées les questions sur la nature du corpus, sur son traitement informatique avec l'objectif d'une exploitation finale. Les limites de ce mémoire ne me permettent pas de développer en profondeur ce sujet du point de vue linguistique mais y a de nombreux travaux à ce sujet. En ce qui concerne les connaissances techniques, les questions sur les normes de codage des documents écrits ou sonores doivent être abordées. On ne peut constituer le corpus sans poser d'emblée la question de son traitement, de son archivage et de son exploitation. Les deux étapes : première (constitution) et dernière (objectif final : consultation, exploitation) sont donc directement liées.

Ces compétences ne font pas souvent partie des formations en TAL ce qui me semble être une erreur car, de mon point de vue, le linguiste taliste est censé être capable de constituer des ressources linguistiques selon les techniques et les normes actuelles. J'observe que les stages en TAL demandent parfois ce niveau de connaissance.

Le traitement des données comprend les différentes méthodes : symboliques et probabilistes. Les étudiants doivent avoir connaissance des deux.

L'analyse des données peut se faire avec des méthodes abordées au cours de l'étape du traitement des corpus mais peut également se faire avec les outils existants. Les étudiants doivent les connaitre aussi et savoir bien les utiliser.

Enfin, la programmation reste indispensable pour assurer l'autonomie des étudiants. Comme je l'ai mentionné, les étudiants linguistes appréhendent ces cours. Notre Master où les cours durent trois semestres ne peut pas entrer en concurrence avec des formations en informatique où les étudiants ont pu aborder des langages de programmation au niveau de la Licence. Nous mettons ainsi de côté les langages d'objet comme Java et ne présentons que des langages de scripts comme Perl ou Python.

En conclusion, l'étudiant doit constituer un corpus selon les méthodes et les techniques actuelles, connaître les formalismes et les modèles utilisés, savoir analyser les résultats obtenus au-delà de leur quantification statistique. J'implique les étudiants dans mes projets de recherches concernant l'annotation et l'analyse des données. Ce travail demande une grande rigueur et des capacités d'observation des variations et des régularités attestées dans le corpus pour assister, perfectionner ou interpréter le processus automatique. Ce sont ces compétences que j'ai essayé de transmettre à mes étudiants de Licence et de Master à travers les cours

« Introduction au TAL, Constitutions de corpus, Outils linguistiques pour l'extraction de l'information, Enrichissement des corpus, Description linguistique pour le TAL, Traitement de l'information etc. »

Dans le Tableau 19, je répertorie l'ensemble des compétences que je juge indispensables pour une formation de linguiste Taliste.

Connaissances passives : connaissances des ressources linguistiques et des outils existants	Lexiques : Delaf, Delac, Lefff, Wordnet, FondamenTAL, JeuDeMots etc. Corpus : FTB, Eslo, CNRTL, Frantext, ANNODIS, CoMeRe, 88milSMS, TCOF, Scientext etc. Outils d'annotation automatique : TreeTagger, Sem, Melt, LiaNE, CasEn, Distagger, Marsatag etc. Concordanciers : AntConc, TXM, anaText etc. Applications en ligne comme « Les Voisins d'En Face » etc.
Constitution des corpus : bruts et annotés	Bien définir la tâche et la nature du corpus corpus écrit : normes XML, TEI etc. corpus oraux : Transcriber, Praat etc. annotation manuelle : Glozz, Gate, ANVIL, ELAN etc. outils d'annotation automatique : TreeTagger, Sem, Melt, LiaNE, CasEn, Distagger etc. Conventions d'annotation : Ester, Quaero, timeMl etc. Langages de balisage et leurs dérivés : XML, XSLT, XPATH, HTML, XHTML Bases de données : SQL, Access, easyPHP etc.
Traitement des corpus et analyse des données	Repérer et extraire une information - méthodes symboliques : Unitex, expressions régulières dans l'éditeur de texte (LibreOffice, Notepad++ etc.) ou sous Linux - feuilles de style XSLT (=>XPATH) Apprentissage automatique : Weka, Le Wapiti etc. Statistiques textuelles et concordanciers : Lexico, AntConc, TXM, Hyperbase etc. Langages de programmations : Perl, Python, Java etc. etc.

Tableau 19 : Synthèse des compétences à acquérir par un linguiste au cours de sa formation en TAL

Si l'on examine ce tableau, ces connaissances peuvent avoir une utilisation plus large. Aujourd'hui, un chercheur dans les sciences sociales et humaines est obligé de constituer un objet d'étude qui est le plus souvent un corpus et il doit être capable de l'analyser avec les outils informatiques existants. Constituer le corpus et savoir l'analyser avec les outils du TAL est donc un atout non négligeable pour un chercheur. C'est la raison du succès actuel que rencontrent les humanités numériques, un domaine où le TAL occupe une place importante. Une autre application possible de ces compétences est le domaine de la représentation des connaissances. A cela, j'ajoute un autre domaine, provenant du monde industriel, où les outils du TAL sont souvent utilisés : la gestion des connaissances « Knowledge management ». Je donne des cours dans les deux branches. Je trouve que la capacité de modéliser l'information linguistique dans le cadre de son annotation peut s'avérer très utile pour pouvoir être capable de représenter les connaissances afin de les traiter ou de les gérer automatiquement. Les avantages sont réciproques. Si l'on connait les méthodes de représentation des connaissances, on peut s'en inspirer pour l'annotation également.

6.5. Mot de la fin

Ce mémoire est un bilan de l'ensemble des activités de recherches que je mène depuis mon doctorat mais ces activités sont loin d'être achevées. Il y a tellement de questions encore à poser et tellement de sujets encore à creuser.

J'ai choisi ce domaine passionnant qui est la langue, cette liberté, cette variation, cette infinité de possibilités qu'elle représente et ce domaine si strict, si rigoureux mais qui est aussi si passionnant – l'informatique. Comment ces deux domaines qui s'opposent par tant d'aspects peuvent se recouper ? Le TAL est la réponse à cette question.

Est-ce que le linguiste se spécialisant dans le TAL peut trouver sa place dans ce domaine ? J'ai essayé de répondre à cette question par ce mémoire. Les connaissances des théories et des modèles linguistiques, la sensibilité particulière à la langue et sa nature, une grande rigueur et des capacités d'observation sont les atouts indéniables. Et même s'il est difficile de prédire l'évolution du TAL, je reste positive et je considère que le linguiste y aura toujours sa place.

Références

Abeillé A., Clément L., Toussenel F. (2003). Building a treebank for French. In A. AbeilléBEILLÉ, (éds.), *Treebanks*. Kluwer, Dordrecht, pp. 165-187.

Abney S. (1996). Partial Parsing via Finite-State Cascades. Proceeding of *the ESSLLI'96 Robust Parsing Workshop*, pp. 8-15.

Abney S. (1991). Parsing by chunks. In R. Berwick, S. Abney, and C. Tenny (eds.), *Principle-based Parsing*. Kluwer Academic Publishers, Dordrecht, pp. 257-278.

Abouda L., Baude O. (2007). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des Eslo. Actes du Colloque *Corpus en Lettres et Sciences sociales des documents numériques à l'interprétation*, Colloque d'Albi Langages et Signification. Presses Universitaires du Mirail, pp. 161-168.

Abouda L., Perrot M.-E. (2006). Une question « embarrassante » en situation d'interview. Modélisation et stratégies de légitimation. *The 3rd Freiburg Workshop on Romance Corpus Linguistics, Corpora and Pragmatics*. Freiburg, Germany.

Adam J.-M. (2001). Entre conseil et consigne : les genres de l'incitation à l'action. *Pratiques*, n° 111/112, pp. 7-38.

Adda-Decker M., Habert B., Barras C., Adda G., Boula De Mareüil P., Paroubek P. (2003). A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech (DiSS'03)*. Gothenburg, University of Gothenburg, pp. 67-70.

Adler S., Eshkol-Taravella I. (2012). « Geste » et « démarche » en tant que noms généraux dans le langage médiatique écrit. *Autour de Pierre Cadiot, Revue de Sémantique et Pragmatique*, n° 31, pp. 90-132.

Adler S., Eshkol-Taravella I. (à paraître). Noms généraux et complexité sémantico-pragmatique. Actes du colloque *Représentations du sens linguistique (RSL VI)*, Nantes, France.

Androutsopoulos I., Malakasiotis P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38, pp. 135–187.

Anis J. (1999). *Internet, communication et langue française.* Paris, Hermès.

Anis J. (2002). Communication électronique scripturale et formes langagières : chats et SMS. Actes des journées *S'écrire avec les outils d'aujourd'hui*. Université de Poitiers.

Anscombre J.-C. (1996). Partitif et localisation temporelle. *Langue française* n°109, pp. 80-103.

Antoine J.-Y., Goulian J., Villaneau J. (2003). Quand le TAL robuste s'attaque au langage parlé : analyse incrémentale pour la compréhension de la parole spontanée. Actes de *TALN2003*, pp. 25-34.

Attal P. (1976). A propos de l'indéfini « des » : problèmes de représentation sémantique. *Le français moderne*, n°64-2, pp. 126-142.

Baccianella S., Esuli A., Sebastiani F. (2010). SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Proceedings of *LREC'10*, pp. 2200-2204.

Badra F., Bendaoud R., Bentebitel R., Champin P., Cojan J., Cordier A., Després S., Jean-Daubias S., Lieber J., Meilender T., Mille A., Nauer E., Napoli A., Toussaint Y. (2008). TAAABLE : Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking. In *ECCBR Workshops, Workshop of the First Computer Cooking Contest*, Springer, Heidelberg, pp. 219–228.

Bar-Hillel Y. (1960). The Present Status of Automatic Translation of Languages. F.C. Alt ed. *Advances in Computers* vol.1, Academic Press, N.Y., London, pp. 91-141.

Baude O., Dugua C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus 10, Varia*, pp. 99-118.

Baude O. (2006). *Corpus oraux : guide des bonnes pratiques*, CNRS-Editions et Presses universitaires d'Orléans.

Béchet F., Charton E. (2010). Unsupervised knowledge acquisition, for extracting named entities from speech. *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP2010)*, pp. 5338-5341.

Béchet F., Sagot B., Stern R. (2011). Coopération de méthodes statistiques et symboliques pour l'adaptation non supervisée d'un système d'étiquetage en entités nommées. Actes de *TALN2011*.

Benzitoun C. (2004). L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ?, *RÉCITAL*, Maroc, Fes.

Benzitoun C., Campione E., Deulofeu J., Henry S., Teston S., Valli A., Véronis J. (2004). L'analyse syntaxique de l'oral : problèmes et méthode. Journée d'étude de l'ATALA *Annotation syntaxique de corpus oraux*, Paris.

Benzitoun C., Fort K., Sagot B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. Actes de *TALN2012*.

Bergounioux G., Eshkol I. (à paraître). Quand faire, c'est dire : l'exemple de la recette. In G. Bergounioux (ed.) *Une étude de cas en linguistique de corpus*, Champion, Paris.

Bhagat R., Hovy E. (2013). What is a paraphrase? *Computational Linguistics*, 39(3), pp. 463–472.

Biber D., Johansson S., Leech G., Conrad S., Finegan E. (1999). *Longman Grammar of Spoken and Written English*, Longman, London.

Bioud M., (2006). *Une normalisation de l'emploi de la majuscule et sa représentation formelle pour un système de vérification automatique des majuscules dans un texte*. Thèse de doctorat, Université de Franche-Comté.

Bird S., Liberman M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33:1-2, pp. 23-60.

Blanc O., Constant M., Dister A., Watrin P. (2008). Corpus oraux et chunking. Actes de *JEP*.

Blanc O., Constant M., Dister A., Watrin P. (2010). Partial parsing of spontaneous spoken french. Proceedings of *LREC'10*.

Blanche-Benveniste C. (1991). Les études sur l'oral et le travail d'écriture de certains poètes contemporains. *Langue française, L'oral dans l'écrit*, vol. 89, pp. 52-71.

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

Blanche-Benveniste C. (2000). Transcription de l'oral et morphologie. In M. Gille, R. Kiesler (eds.) *Romania Una et diversa, Philologische Studien für Theodor Berchem*. Gunter Narr, Tübingen, pp. 61-74.

Blanche-Benveniste C., Bilger, M., Rouget C., Eynde K. (1990). *Le français parlé – études grammaticales*. Sciences du Langage. CNRS Editions, Paris.

Blanche-Benveniste C., Jeanjean C. (1987). *Le français parlé. Transcription et édition*. Institut national de la Langue française. Didier Érudition, Paris.

Blanche-Benveniste C. (2005). Les aspects dynamiques de la composition sémantique de l'oral. In A. Condamines (dir.) *Sémantique et corpus*, Hermès, Londres, pp. 40-73.

Boons J.-P. (1987). La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs. *Langue Française*, n° 76, pp. 5-40.

Borillo A. (1998). *L'espace et son expression en français.* L'essentiel, Ed. Orphrys, Paris.

Bouamor H. (2009). Extraction des connaissances à partir du Web pour la recherche des images géoréférencées. *CORIA*, pp. 519-526.

Bouamor H. (2012). *Étude de la paraphrase sous-phrastique en traitement automatique des langues*. Thèse de doctorat, Université Paris Sud, Paris.

Bouraoui J.-L., Canitrot M. (2013). FMO : outil d'analyse automatique de l'opinion. Actes de *TALN2013*.

Bradley M. M., Lang P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Campione E., Véronis J. (2004). Pauses et hésitations en français spontané. *JEP'2004*, pp. 109-112.

Candéa M. (2000). Les euh et les allongements dits d'« hésitation » : deux phénomènes soumis à certaines contraintes en français oral non lu. *JEP'2000,* pp. 73-76.

Cappeau P., Gadet F. (2007). Où en sont les corpus sur les français parlés ? *Revue Française de Linguistique Appliquée*, vol. XII, pp. 129-133.

Cappeau P., Schnedecker C. (2014). Gens, personne(s), individu(s) : trois saisies de l'humain. *Actes du CMLF 2014*, pp. 3027-3040.

Carletta J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, n˚22, pp. 249–254.

Chanet C. (2001). Connecteurs, particules et représentations cognitives de la planification discursive. In E.T. Nemeth (ed.), *Cognition in language use: Selected papers from the 7th International Pragmatics Conference,* vol. 1, International Pragmatics Association, Antwerp, pp. 44-55.

Chanet C. (2004). Fréquence des marqueurs discursifs en français parlé : quelques problèmes de méthodologie. *Recherches sur le français parlé*, n°18, pp. 83-106.

Charaudeau P. (1983). *Langage et Discours - Eléments de sémiolinguistique*, Hachette, Paris.

Charaudeau P., Maingueneau D. (2002). *Dictionnaire d'analyse du discours*. Éditions du Seuil, Paris.

Chinchor N. (1997). Muc-7 Named Entity Task Definition. *MUC-7*, Fairfax, Virginia.

Christodoulides G., Avanzi M., Goldman J.-P. (2014). DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech. Proceedings of *LREC'14.*

Christodoulides G., Grosman I. (2012). DisMo. Un outil d'annotation morphosyntaxique pour le français parlé. Journée d'études CONSCILA, *Annotation syntaxique de corpus oraux*. ENS Paris.

Clédat L. (1901). La préposition et l'article partitif. *Revue de philologie française et de littérature*, XV, pp. 81-131.

Clément L., Sagot B., Lang B. (2004). Morphology based automatic acquisition of large-coverage lexical. Proceedings of *LREC2004*, pp. 1841-1844.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), pp. 37–46.

Colson M. (2012). La recette du *blanc-manger* : de la variation à la dégustation. Actes de *4e Journée liégeoise de Traitement des Sources galloromanes (TraSoGal).*

Constant M., Tellier I. (2012). Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. Proceedings of *LREC'12,* pp. 646-650.

Constant M., Tellier I. (2013). Intégrer des ressources lexicales et grammaticales externes dans des analyseurs partiels probabilistes. Actes de *TALN2013.*

Constant M., Tellier I., Duchier D., Dupont Y., Sigogne A., Billot S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. Actes de *TALN2011*.

Corblin F. (1983). Les désignateurs dans les romans. *Poétique*, 54, pp. 199-211.

Corblin F., Gardent C. (2005). Contexte et interprétation. In F.Corblin & C.Gardent (Eds.), *Interpréter en contexte*, Hermès Science Publication, Paris, pp. 15-28.

Cori M., David S., Léon J. (2008). Eléments de réflexion sur la place des corpus en linguistique. *Langages*, pp. 5-11.

Cori M., Léon J. (2002). La constitution du TAL. Etude historique des dénominations et des concepts. *TAL*, vol. 43, n˚3, pp. 21-55.

Cosnier J. (1988). Grands tours et petits tours. In J. Cosnier, N. Gelas & C. Kerbrat-Orecchiono (Eds.), *Echanges sur la conversation*, Editions du CNRS, Lyon, pp. 175-184.

Courtois B. (2009). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, vol. 87, pp. 11-22.

Courtois B., Garrigues M., Gross G., Gross M., Jung R., Mathieu-Colas M., Monceaux A., Poncet-Montange A., Silberztein M., Vivés R. (1997). *Dictionnaire électronique DELAC : les mots composés binaires*, Technical Report n°56, University Paris 7, LADL.

Courtois B., Silberztein M. (1990). Dictionnaires électroniques du français, *Langue française*, 87, pp. 11-22.

Crabbé B., Candito M. (2008). Expériences d'analyse syntaxique du français. Actes de *TALN2008*.

Culioli A. (1976). *Notes du séminaire de DEA, 1983-84*. Paris.

Cyril G., Zweigenbaum P., Paroubek P. (2013). DEFT2013 se met à table : présentation du défi et résultats. Actes de *TALN2013*.

Dauzet A. (1963). *Dictionnaire étymologique des noms de lieux en France*. Larousse, Paris.

De Gaulmyn M.-M. (1987). Les régulateurs verbaux : le contrôle des récepteurs. In J. Cosnier & C. Kerbrat-Orecchiono (Eds.), *Décrire la Conversation*, Presses Universitaires de Lyon, Lyon, pp. 203-223.

Debrock M., Mertens P., Truyen F., Brosens V. (2000). *ELICOP, Etude LInguistique de la COmmunication Parlée : Constitution et exploitation d'un corpus de français parlé automatisé*. Departement Linguïstiek, K.U.Leuven.

Denis P., Sagot B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for stateof-the-art POS tagging with less human effort. Proceedings of *Pacific Asia Conference on Language (PACLIC 2009)*, Information and Computation, Hong Kong, China.

Denis P., Sagot B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. Actes de *TALN2010*.

Dister A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelle orale VALIBEL*. Thèse de doctorat, Université de Louvain.

Domingues C., Eshkol-Taravella I. (2013). Repérer des toponymes dans des titres de cartes topographiques. Actes de *TALN2013*, (article court).

Domingues C., Eshkol-Taravella I. (à paraître a). Ecriture des toponymes en français: variations entre normes et usage. Actes du colloque international *Normes linguistiques et textuelles : émergence, variations, conflits*, 26-27 mars 2015, Toulon, France.

Domingues C., Eshkol-Taravella I. (à paraître b). Toponym recognition in custom-made map titles. *International Journal of Cartography*, Taylor & Francis.

Dubois J. (1994). *Dictionnaire de linguistique et des sciences du langage*. Larousse, Paris.

Dufour-Lussier V., Le Ber F., Lieber J. (2011). Quels formalismes temporels pour représenter des connaissances extraites de textes de recettes de cuisine ? *6e atelier de représentation et raisonnement sur le temps et l'espace (RTE 2011)*, Chambéry, France, pp. 16–25.

Dufour-Lussier V., Lieber J., Nauer E., Toussaint Y. (2010). Analyse formelle de concepts pour l'adaptation de cas textuels. *18e atelier de raisonnement à partir de cas*, Strasbourg, France, pp. 71–82.

Ehrmann M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. Thèse de doctorat, Université Paris 7 - Centre de recherche Xerox, Grenoble (XRCE).

Engel L. (2000). Syntaxe à la une : la structure des titres de journaux français et britanniques. *The Web Journal of French Media Studies 3/1*.

Englebert A. (1996). L'article partitif : l'évolution des conditions d'emploi. *Langue française*, 109, pp. 9-28

Eshkol I. (2015). Interpréter le contexte dans un corpus oral : fonctions et limites du traitement automatique des données linguistiques. In J.-P. Miller (ed.) *Le contexte - Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels.* Editions Chemins de tr@verse sur Bouquineo.fr, Paris, pp. 67-80.

Eshkol I. (2010)**.** Entrer dans l'anonymat. Etude des "entités dénommantes" dans un corpus oral. In N. Pepin & E. De Stefani (eds.) *Eigennamen in der gesprochenen Sprache*, Francke VERLAG, pp. 245-266.

Eshkol I., Tellier I, Taalab S., Billot S. (2010). Etiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques, *10es Journées Internationales d'analyse statistique des données textuelles JADT 2010*.

Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C., Tellier I. (2012). Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012. *TAL : Ressources linguistiques libres*, vol. 52, n° 3, pp. 17-46.

Eshkol-Taravella I., Grabar N. (2014a). Paraphrastic reformulations in spoken corpora. *Advances in Natural Language Processing Lecture Notes in Computer Science.* Proceedings of *9th International Conference on NLP, PolTAL 2014*, Vol. 8686, Springer, pp. 425-437.

Eshkol-Taravella I., Grabar N. (2014b). Repérage et analyse de la reformulation paraphrastique dans les corpus oraux. Actes de *TALN2014*.

Eshkol-Taravella I., Grabar N. (à paraître). Approximation à travers la reformulation paraphrastique. Actes du colloque *Cerlico2015 : Linéarité et interprétation 2. Approximation, modulation, ajustement*, 11-12 juin 2015, Rennes, France.

Eshkol-Taravella I., Baude O., Maurel D., Kanaan-Caillol L. (2015). Recherche des indices permettant une identification: l'anonymisation des transcriptions du corpus ESLO. In Actes de *TALN2015* (atelier ETeRNAL).

Esuli A., Sebastiani F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. Proceedings of *LREC'06*, vol. 6, pp. 417-422.

Fairon C., Kein J.-R. (2010). Les écritures et graphies inventives des SMS face aux graphies normées. *Le français d'aujourd'hui*. *Graphies : signes, gestes, supports*, 170, pp. 113-122.

Fairon C., Kein J.-R., Paumier S. (2007). *Le langage SMS : étude d'un corpus informatisé à partir de l'enquête « Faites don de vos SMS à la science »*. Cahiers du CENTAL 3.1, Presses Universitaires de Louvain, Louvain-la-Neuve.

Fellbaum C. (2005). WordNet and wordnets. In K.Brown *et al*. (eds.), *Encyclopedia of Language and Linguistics*. Second Edition, vol.13, Elsevier, Oxford, pp. 665-670.

Fernandez M.-J. (1994). *Les particules énonciatives*. PUF, Paris.

Francis G. (1994). Labelling discourse: an aspect of nominal-group lexical cohesion. In Coulthard M. (ed). *Advances in written text analysis*, London, Routledge, pp. 83-101.

Flottum K. (1995). *Dire et redire. La reformulation introduite par « c'est-à-dire »*. Thèse de doctorat, Hogskolen i Stavanger, Stavanger.

Friburger N. (2002). *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*. Thèse de doctorat, Université François Rabelais de Tours.

Friburger N., Maurel D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, vol. 313, pp. 94-104.

Fuchs C. (1982). *La paraphrase*. Paris : PUF.

Fuchs C. (1994). *Paraphrase et énonciation*. Orphys, Paris.

Furet C. (1995) *Le titre. Pour donner envie de lire*. Centre de Formation et de Perfection, Paris.

Gadet F. (1992). *Le Français populaire*. Paris, Presses universitaires de France (coll. « *Que sais-je ?* »).

Gadet F. (2003). *La variation sociale en français*. Ophrys, Paris.

Gaio M., Sallaberry C., Nguyen V. T. (2012). Typage de noms toponymiques à des fins d'indexation géographique. *TAL*, vol. 53, n°2, pp. 143-176.

Galliano S., Gravier G., Chaubard L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts, *Interspeech 2009*.

Galmiche M. (1986). Note sur les noms de masse et le partitif. *Langue française*, vol.72, n°1, pp. 40-53.

Gazeau M. A., Maurel D. (2006). Un dictionnaire INTEX de noms de professions : quels féminins possibles ? *Cahiers de la MSH Ledoux*, pp. 115-127.

Gouvert X. (2008). *Problèmes et méthodes en toponymie française. Essais de linguistique historique sur les noms de lieux du Roannais.* Thèse de doctorat. Université de Paris Sorbonne (Paris IV).

Grabar N., Eshkol-Taravella I. (2015). ...des conférences enfin disons des causeries... Détection automatique de segments en relation de paraphrase dans les reformulations de corpus oraux. Actes de *TALN2015*.

Greimas A. (1983). *Du Sens II, essais sémiotiques*. Paris, Le Seuil.

Grevisse M., Goosse A. (1993). *Le Bon Usage*. Duculot. Paris, Louvain-la-Neuve.

Grosjean F., Deschamps A. (1972). Analyse des variables temporelles du français spontané. *Phonetica*, 26, pp. 129-156.

Grosjean F., Deschamps A. (1973). Analyse des variables temporelles du français spontané. II. Comparaison du français oral dans la description avec l'anglais (description) et avec le français (interview radiophonique). *Phonetica*, 28, pp. 191-226.

Grosjean F., Deschamps A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica*, 31, pp. 144-184.

Gross G. (1996). *Les expressions figées en français. Noms composés et autres locutions*. Orphrys, Paris.

Gross M. (1982). Une classification des phrases « figées » du français. *Revue québécoise de linguistique*, vol. 11, n°2, pp. 151-185.

Gross M. (1997). The construction of local grammars. In D. J. Lipcoll, D. H. Lawrie & A. H. Sameh (Eds.), *Finite-State Language Processing*, The MIT Press, Cambridge, Mass, pp. 329–352.

Grouin C., Zweigenbaum P. (2011). Une approche à plusieurs étapes pour anonymiser des documents médicaux. *RSTI-RIA*, 25:4, pp. 525-549.

Guimier C. (2002). *Les adverbes du français*. Ophrys, Paris

Gulich E., Kotschi T. (1983). Les marqueurs de la reformulation paraphrastique. *Cahiers de linguistique française*, 5, pp. 305–351.

Gülich E., Kotschi T. (1987). Les actes de reformulation dans la consultation La dame de Caluire. In P. Bange, (ed.), *L'analyse des interactions verbales. La dame de Caluire : une consultation*, P Lang, Berne, pp. 15–81.

Habert B. (2001). Des corpus représentatifs : de quoi, pour quoi, comment ? In M. Bilger (ed.), *Linguistique sur corpus. Études et réflexions*, Presses Universitaires de Perpignan, Perpignan, pp. 11-58.

Habert B. (2004). Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs. *Revue Française de Linguistique Appliquée*, vol.9, n°1, pp. 5-24.

Habert B. (2005). Portrait de linguiste(s) à l'instrument. *Texto*, vol.10, n°4, pp. 124-132

Habert B. (2006). TAL sur corpus : histoire, acquis, défis. In. G.Sabah (ed.) *Compréhension des langues et interaction*, Lavoisier, Hermes, Paris, pp. 249-271.

Habert B., Nazarenko A. (1997). *Les linguistiques de corpus*. A. Colin, Paris.

Hagège C. (1985). *L'homme de paroles. Contribution linguistique aux sciences humaines*. Fayard, Paris.

Hamon P. (1977). Pour un statut sémiologique du personnage. In R. Barthes *et al.* (éds.), *Poétique du récit*, Points-Seuil, Paris.

Hatzivassiloglou V., McKeown K. (1997). Predicting the semantic orientation of adjectives. In Proceedings of *The eighth conference on European chapter of the Association for Computational Linguistics*, Association for Computational Linguistics (ACL), Morristown, NJ, USA, pp.174-181.

Henry S. (2005). Quelles répétitions à l'oral ? Esquisse d'une typologie, G. Williams (Éd.), *La Linguistique de corpus*. Presses universitaires de Rennes, Rennes, pp. 81-92.

Heurley L. (1997). Vers une définition du concept de texte procédural : le point de vue de la psycholinguistique. *Les Cahiers du Français Contemporain*, 4, pp. 109-133.

Heurley L. (2001). Compréhension et utilisation de textes procéduraux : l'effet de l'ordre de mention des informations. *Revue Française de Linguistique Appliquée 2001/2*, vol. VI, pp. 29-46.

Heurley L., Ganier F. (2006). L'utilisation des textes procéduraux : Lecture, compréhension et exécution d'instructions écrites. *Intellectica*, 44/2, pp. 45-62.

Hoek L. H. (2004). *Titre, toiles et critique d'art : déterminants institutionnels du discours sur l'art au XIXe siècle*. Ed. Rodopi B.V, Amsterdam.

Halliday M.A.K., Hasan R. (1976). *Cohesion in English.* Longman, London.

Honeycutt C., Herring S. (2009). Beyond Microblogging: Conversation and Collaboration in Twitter. Proceedings of *42nd Hawaii International Conference on System Science (HICSS)*, IEEE Press, Washington, pp. 1-10.

Houdebine-Gravaud A.-M. (2003). *L'imaginaire linguistique.* L'Harmattan, Paris.

Huet S., Gravier G., Sébillot P. (2006). Peut-on utiliser les étiqueteurs morphosyntaxiques pour améliorer la transcription automatique. *26èmes Journées d'Études sur la Parole (JEP)*, Dinard, France.

Ide N. (2007). Annotation Science: From Theory to Practice and Use. (Invited Talk) Data Structures for Linguistics Resources and Applications. Proceedings of *the Biennial GLDV Conference*.

Ide N., Romary L. (2004). A Registry of Standard Data Categories for Linguistic Annotation. Proceedings of *LREC2004*, pp. 135-39.

Jackiewicz A. (2014). Études sur l'évaluation axiologique : présentation. *Langue française 4*, n° 184, pp. 5-16.

Jacques M.-P., Poibeau T. (2010). Étudier des structures de discours : préoccupations pratiques et méthodologiques. *Corela*, vol.8, n° 2.

Java A., Song X., Finn T., Tseng B. (2007). Why we Twitter: Understanding microblogging usage and communities. In Proceedings of *Joint 9th WEBKDD and 1st SNA-KDD* Workshop, ACM Press.

Jonasson K. (1994). *Le Nom propre. Constructions et interprétations*. Duculot, Louvain-la-Neuve, Belgique.

Kanaan L. (2011). *Reformulations, contacts de langues et compétence de communication : analyse linguistique et interactionnelle dans des discussions entre jeunes Libanais francophones*. Thèse de doctorat, Université d'Orléans, Orléans.

Kerbrat-Orecchioni C. (1999). *L'Énonciation. De la subjectivité dans le langage.* Armand Colin, Paris.

Kleiber G. (1981). *Problèmes de référence : descriptions définies et noms propres*. Paris, Klincksieck.

Kleiber G. (1984). Dénomination et relations dénominatives. *Langages*, 76, pp. 77-94.

Kolhatkar V., Zinsmeister H., Hirst G. (2013a). Annotating Anaphoric Shell Nouns with their Antecedents. Proceedings of *the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 112–121.

Kolhatkar V., Zinsmeister H., Hirst G. (2013b). Interpreting Anaphoric Shell Nouns using Antecedents of Cataphoric Shell Nouns as Training Data. Proceedings of *the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 300–310.

Kripke S. A. (1972). Naming and Necessity. In D. Davidson, G. Harman (éds.). *Semantics of Natural Language*, Dordrecht : Reidel, pp. 253-355.

Kupferman L. (1994). *Du* : un autre indéfini ? *Faits de langue*, 4, pp. 195-203.

Kupferman L. (1996). Les génitifs : gouvernement d'antécédent et gouvernement thématique, *Langue française*, 109, pp. 104-125.

Kupferman L. (1998). *Des* : pluriel de *du* ?. In M. Bilger, K. van den Eynde, F. Gadet (sld.), *Analyse linguistique et approches de l'oral. Recueil d'études offert en hommage à Claire Blanche-Benveniste*, Peeters, Louvain/Paris, pp. 229-238.

Kupferman L. (1999). Réflexions sur la partition : les groupes nominaux partitifs et la relativisation. *Langue française*, 122, pp. 30-51.

Kupferman L. (2001). Quantification et détermination dans les groupes nominaux. In X. Blanco, P.-A. Buvet, Z. Gavrilidou (sld.), *Détermination et formalisation*, John Benjamins, Amsterdam-Philadelphie, pp. 219-234.

Lafferty J., McCallum A., Pereira F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Proceedings of *ICML'01*, pp. 282–289.

Laporte E., Monceaux A. (1999). Elimination of lexical ambiguities by grammars : the ELAG system. In C. Fairon (Éd.), *Analyse lexicale et syntaxique : le système INTEX, Lingvisticæ Investigationes 22*, (1-2), John Benjamins, Amsterdam-Philadelphia, pp. 341-367.

Laur D. (1991). *Sémantique du déplacement et de la localisation en français : une étude des verbes, des prépositions et de leur relation dans la phrase simple*. Thèse de doctorat, Université de Toulouse I.

Laurent D., Nègre S., Séguéla P. (2009a). Apport des cooccurrences à la correction et à l'analyse syntaxique. Actes de *TALN2009*.

Laurent D., Nègre S., Séguéla P. (2009b). L'analyseur syntaxique Cordial dans Passage. Actes de *TALN2009.*

Le Pesant D. (2011a). Problèmes de morphologie, de syntaxe et de classification sémantique dans le domaine des prépositions locatives. In F. Neveu, P. Blumenthal et N. Le Querler (dir.), *Au commencement était le verbe. Syntaxe, Sémantique et Cognition. Mélanges en l'honneur du Professeur Jacques François*. Peter Lang, Bern, Berlin, pp. 349-372.

Le Pesant D. (2011b). Vers un thesaurus syntactico-sémantique des mots d'affect. *Cahiers de Lexicologie 2*, n° 99. In M. Fasciolo (éd.). *Lexique et philosophie*, Classiques Garnier, Paris, pp. 117-132.

Le Pesant D. (2012). Essai de classification des prépositions de localisation. Actes du *CMLF 2012*, pp. 921-936.

Leech G. (1994). 100 million words of English : the British National Corpus. *English Today*, 9(1), pp. 9-15.

Leech G. (1997). Introduction corpus annotation. In R. Garside, G. Leech, A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora*. London, Longman, pp. 1-18.

Leech G. (1991). The state of the art in corpus linguistics. In K. Aijmer, B. Altenberg (eds.), *English Corpus Linguistics.* Longman, London, pp. 8-29.

Legallois D. (2008). Sur quelques caractéristiques des noms sous-spécifiés. *Scolia*, n° 23, pp. 109–127.

Leroy S. (2004). *Le nom propre en français.* Ophrys, Paris.

Lesbegueries J. (2007). *Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialise*. Thèse de doctorat, Université de Pau et des Pays de l'Adour.

Longhi J. (2012). Discours, style, format : niveaux de structuration de la textualité des Tweets de Mouloud. Actes du *CMLF 2012*, pp. 1127-1141.

Longhi J. (2013). Essai de caractérisation du tweet politique. *L'Information Grammaticale*, n°136, pp. 25-32.

Lopez C. (2012). *Titrage automatique de documents textuels*. Thèse de doctorat, Université de Montpellier 2.

Lorenz P. (2013). *Le chat en tant que phénomène langagier : étude comparative français-espagnol-polonais*. PAF.

Madnani N., Dorr B. J. (2010). Generating phrasal and sentential paraphrases : A survey of data-driven methods. *Computational Linguistics*, 36, pp. 341–387.

Mahlberg M. (2005). *English general nouns; a corpus theoretical approach*. John Benjamins company, Amsterdam.

Marchand M., Mesnard O., Besançon R., Vilnat A. (2014). Influence des marqueurs multi-polaires dépendant du domaine pour la fouille d'opinion au niveau du texte. Actes de *TALN2014*.

Marcus M., Santorini B., Marcinkiewicz M. A. (1993). Building a large annotated corpus of english : The penn treebank. *Computational Linguistics*, 19(2), pp. 313–330.

Mårdh I. (1980). *Headlinese : on the grammar of English front page headlines (Lund Studies in English 58)*, CWK Gleerup, Lund.

Martin R. (1976). *Inférence, antonymie et paraphrase*. Klincksieck, Paris.

Mathieu Y. (1999). Les prédicats de sentiment. *Langages*, vol. 33, n° 136, pp. 41-52.

Mathieu-Colas M. (1998). La majuscule flottante. Remarques sur l'orthographe des noms propres composés (type *N Adj*). *BULAG* n° 23, Centre Lucien Tesnière, Université de Franche-Comté, Besançon, pp. 123-144.

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., Nouvel D. (2011). Cascades de trans-ducteurs autour de la reconnaissance des entités nommées. *TAL*, vol. 52, n° 1, pp. 69-96.

Maurel D., Friburger N., 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, pp. 220-223.

Mauro G, Sallaberry C., Nguyen V. T. (2012). Typage de noms toponymiques à des fins d'indexation géographique. *TAL*, vol. 53, p. 1-35.

McCallum A., Li W. (2003). Early results for named entity recognition with conditional random fields. Proceedings of *CoNLL 2003*.

Mélanie-Becquet F., Landragin F. (2014). Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques. *Langages 3*, n° 195, pp. 117-137.

Melčuk I. (1988). Paraphrase et lexique dans la théorie linguistique sens-texte. Vingt ans après. *Cahiers de lexicologie*, vol. 52, n°1, pp. 5-50.

Mertens P. (2002). Les corpus de français parlé ELICOP : consultation et exploitation. In J. Binon et al. (éd.), *Tableaux Vivants. Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock*. Universitaire Pers, Leuven.

Meystre S., Friedlin B S., Shuying S., Samore M. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, vol. 10, n°1.

Moirand S. (1975). Le rôle anaphorique de la nominalisation dans la presse écrite. *Langue Française*, 28, pp. 60-77.

Morel M.-A., Danon-Boileau L. (1998). *Grammaire de l'intonation, l'exemple du français*, Ophrys, Paris.

Mouillaud M. (1982). Grammaire et idéologie du titre de journal. *Mots*, 4, pp. 69-91.

Muzerelle J., Lefeuvre A., Schang E., Antoine J.-Y., Pelletier A., Maurel D., Eshkol I., Villaneau J. (2014). ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. Proceedings of *LREC'14*.

Muzerelle J., Schang E., Antoine J.-Y., Eshkol I., Maurel D., Boyer A., Novel D. (2012). Annotations en chaînes de coréférences et anaphores dans un corpus de discours spontané en français. *Actes du CMLF 2012*, pp. 2497 - 2516.

Nadeau N., Sekine S. (2009). *A survey of named entity recognition and classification*. In S. Sekine & E. Ranchhod (eds.), John Benjamins publishing company, Amsterdam, pp. 3-28.

Nazarenko A. (2006). Le point sur l'état actuel des connaissances en TAL. In G.Sabah (ed.) *Compréhension des langues et interaction*, Lavoisier, Hermès, Paris, pp. 31-70.

Pak A., Paroubek P. (2010). Le microblogage pour la microanalyse des sentiments et opinions. *TAL 51*, n° 3, pp. 75-100.

Pallaud B. (2002). Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le français parlé*, 17, Université de Provence, pp. 79-101.

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2013). Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS. *Épistémè - revue internationale de sciences sociales appliquées, 9 : Des usages numériques aux pratiques scripturales électroniques*, pp. 107-138.

Panckhurst R., Moïse C. (2014). French text messages. From SMS data collection to preliminary analysis. In L.-A. Cougnon, C. Fairon (éds.), *SMS Communication. A Linguistic Approach*. John Benjamins : Amsterdam-Philadelphia, pp. 141-168.

Paroubek P., Vilnat A., Robba I. (2007). Les résultats de la campagne easy d'évaluation des analyseurs syntaxiques du français. Actes de *TALN2007*.

Pasça M., Dienes P. (2005). Aligning needles in a haystack : Paraphrase acquisition across the Web. Proceedings of *the Second international joint conference on Natural Language Processing (IJCNLP)*, Springer-Verlag, Berlin, Heidelberg, pp. 119–130.

Paumier S. (2003). *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique*, Thèse de doctorat, Université de Marne-la-Vallée.

Picoche J. (1986). *Structures sémantiques du lexique français*. Nathan, Paris.

Plantin C. (2011). *Les bonnes raisons des émotions. Principes et méthode pour l'étude du discours émotionné*. Peter Lang, Berne.

Poitou J. (2009). *Règles typographiques, codes et usages.* http://j.poitou.free.fr/pro/html/typ/codes.html [consulté le 02/10/2012].

Quirk C., Brockett C., Dolan W. (2004). Monolingual machine translation for paraphrase generation. Proceedings of *EMNLP*, Barcelona, pp. 142–149.

Raaj N. (2012). *Automated Tool for Anonymization of Patient Records.* Report. MSc Computing and Management, Imperiel College, London.

Raingeard M., Lorscheider U. (1977). Édition d'un corpus de français parlé. *Recherches sur le français parlé*, 1, pp. 14-29.

Riegel M., Pellat J.-C., Rioul R. (1994). *Grammaire Méthodique du Français*. PUF, Paris.

Rossari C. (1990). Projet pour une typologie des opérations de reformulation. *Cahiers de linguistique française*, 11, pp. 345–359.

Rossari C. (1992). De l'exploitation de quelques connecteurs reformulatifs dans la gestion des articulations discursives. *Pratiques*, 75, pp. 111–124.

Rossari C. (1994). Les opérations de reformulation. Analyse du processus et des marques dans une perspective contrastive français-italien, Peter Lang, Berne.

Rosset S., Grouin C., Zweigenbaum P. (2011). *Entités nommées structurées : guide d'annotation Quaero.* Notes et documents LIMSI N°2011-04.

Roulet E. (1987). Complétude interactive et connecteurs reformulatifs. *Cahiers de linguistique française*, 8, pp. 111–140.

Roulet E., Auchlin A., Moeschler J., Schelling M., Rubattel C. (1985). *L'articulation du discours en français contemporain*. Lang, Bern.

Roze C., Charnois T., Legallois D., Ferrari S., Salles M. (2014). Identification des noms sous-spécifiés, signaux de l'organisation discursive. Actes de *TALN2014.*

Sagot B. (2010). The Lefff, a freely available, accurate and large -coverage lexicon for French. Proceedings of *LREC'10*.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. Proceedings of *International Conference on New Methods in Language Processing*, Manchester, UK.

Schmid H.-J. (2000). *English Abstract Nouns as Conceptual shells*. Berlin, New York, Mouton de Gruyter.

Schnedecker C. (2011). Monsieur tout le monde, Maud Machin-Chouette, Denise Trucmuche, et les autres.... Inventaire et comportement des noms propres "indéfinis" du français. In D. A. Miot, W. De Mulder, E. Moline & D. Stosic (éds.). *Ars Grammatica. Hommage à Nelly Flaux*, Peter Lang, Berne, pp.37-54.

Schnedecker C. (2015). Un problème à la croisée des disciplines linguistiques : les noms d'humains comme interface entre morphologie, syntaxe et sémantique. In A. Rabatel, A. Ferrara-Léturgie & A. Léturgie (éds.), *La sémantique et ses interfaces. Actes du colloque 2013 de l'ASL (Association des sciences du langage)*, pp. 111-141.

Seretan V., Nerima L., Wehrli E. (2003). Extraction of multi-word collocations using syntactic bigram composition. Proceedings of *RANLP2003*, pp. 424–431.

Sha F., Pereira F. (2003). Shallow parsing with conditional random fields. Proceedings of *HLT-NAACL 2003*, pp. 213-220.

Sha F., Pereira F. (2003). Shallow parsing with conditional random fields. Proceedings of *HLT-NAACL*.

Siblot P. (2007). Nomination et point de vue : la composante déictique des catégorisations lexicales. In G.Cislaru, O.Guérin, K.Morin, E.Nee, T.Pagnier & M.Véniard (Eds.). *L'acte de nommer. Une dynamique entre langue et discours*. Paris, Presses Sorbonne Nouvelle, pp. 25-38.

Silberztein M. (2000). Intex : an fst toolbox. *Theoretical Computer Science*, 231(1), pp. 33-46.

Sinclair J. (1996). *Preliminary recommendations on Corpus Typology*. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).

Sinclair J., Wynne M. (2005). *Developing Linguistic Corpora: a Guide to Good Practice Corpus and Text - Basic Principles*. Oxford: Oxbow Books, pp. 1-16.

Sullet-Nylander F. (1998). *Le titre de presse - analyses syntaxique, pragmatique et rhétorique*. Stockholm : Akademitryk AB.

Sutton C., McCallum A. (2006). An Introduction to Conditional Random Fields for Relational Learning. In L. Getoor & B. Taskar (éds.), *Introduction to Statistical Relational Learning*. MIT Press, lise getoor and ben taskar édition, chapitre 4, pp. 93-128.

Tellier I., Duchier D., Eshkol I., Courmet A., Martinet M. (2012). Apprentissage automatique d'un chunker pour le français. Actes de *TALN2012* (article court).

Tellier I., Dupont Y., Courmet A. (2012). Un segmenteur-étiqueteur et un chunker pour le français. Actes de *TALN2012*, (session démo).

Tellier I., Dupont Y., Eshkol I., Wang I. (2013). Adapt a Text-Oriented Chunker for Oral Data: How Much Manual Effort is Necessary? *The 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2013)*, Special Session on Text Data Learning, Hefei, China.

Tellier I., Eshkol-Taravella I., Dupont Y., Wang I. (2014). Peut-on bien chunker avec de mauvaises étiquettes POS ? Actes de *TALN2014*.

Tellier I., Eshkol I., Taalab S., Prost J-P. (2010). POS-tagging for Oral Texts with CRF and Category Decomposition. *Research in Computer Science, special issue : Natural Language Processing and its Applications*, vol. 46, pp. 79-90.

Tellier I., Tommasi M. (2011). Champs Markoviens Conditionnels pour l'extraction d'information. *Modèles probabilistes pour l'accès à l'information textuelle*, Hermès, Paris, pp. 223-267.

Tran M., Maurel D., (2006). Prolexbase : un dictionnaire relationnel multilingue de noms propres, *TAL*, vol. 47, n° 3, pp. 115-139.

Tveit A., Edsberg O., Brox Røst T., Faxvaag A., Nytrø Ø., Nordgård T., Thorsen Ranang T., Grimsmo A. (2004). Anonymization of General Practitioner Medical Records. *Second HelsIT Conference at the Healthcare Informatics*, Trondheim, Norway.

Uzuner O., Luo Y., Szolovits P. (2007). *Evaluating the state-of-the-art in automatic de-identification*. J Am Med Inform Assoc, 14, pp. 550-563.

Valli A., Véronis J. (1999). Etiquetage grammatical des corpus de parole : problèmes et perspectives. L'oral spontané. *Revue Française de Linguistique Appliquée*, vol. IV-2, pp. 113-133.

Van Dijk Tean A. (1985). *Structures of news in the press*. In A.Tean van Dijk (éd), *Discourse and Communication New Approaches to the Analysis of Mass Media Discourse and Communication.* Berlin/New-York: Mouton de Gruyter, pp. 69-93.

Vandeloise C. (1986). *L'espace en français.* Ed. Seuil, Paris, France.

Vanderveken D. (1988). *Les actes de discours : essai de philosophie du langage et de l'esprit sur la signification des énonciations.* Pierre Mardaga éditeur. Bruxelles.

Vernier M. (2011). *Analyse à granularité fine de la subjectivité*. Thèse de doctorat, Université de Nantes.

Véronis J., Guimier de Neef E. (2006). Le traitement des nouvelles formes de communication écrite. In G. Sabah(éd.), *Compréhension automatique des langues et interaction.* Hermès, Paris.

Vezin L. (1976). Les paraphrases : étude sémantique, leur rôle dans l'apprentissage. *L'année psychologique*, 76(1), pp. 177–197.

Vieu L. (1991). *Sémantique des relations spatiales et inférences spatio-temporelles: une contribution à l'étude des structures formelles de l'espace Toulouse*. Thèse de doctorat, Université Paul Sabatier.

Vila M., Antònia Mart M., Rodríguez H. (2011). Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, vol.46, pp. 83–90.

Virbel J. (1997). Contributions de la théorie des actes de discours à une taxinomie des consignes. *In* J. Virbel, J.-M. Cellier & J.-L. Nespoulous (éds.), *Cognition, Discours Procédural, Action,* vol. II. Actes de l'Atelier « Texte et Communication », Prescot, Toulouse, pp. 1-44.

Wagner R. A., Fischer M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1), pp. 168-173.

Wisniewski G., Max A., Yvon F. (2010). Recueil et analyse d'un corpus écologique de corrections orthographiques extrait des révisions de Wikipedia. Actes de *TALN2010*.

Rapports

CNI-CNIG (2010), Recommandations et observations grammaticales. http://www.cnig.gouv.fr/Front/docs/cms/cnt-grammaire-recommandation_126924688421947500.pdf [consulté le 18/12/2012].

Les spécifications d'écriture des toponymes à l'IGN (document confidentiel), http://sbv.ign.fr/outils/pdf/SBV-MAJEC-IT-056-equipement-toponymie.pdf [consulté le 02/10/2012].

Annexes

Annexe 1 : Liste des étiquettes de Cordial Analyseur

ADJFP Adjectif Féminin Pluriel
ADJFS Adjectif Féminin Singulier
ADJHFS Adjectif Féminin Singulier débutant par un "h" aspiré
ADJHMS Adjectif Masculin Singulier débutant par un "h" aspiré
ADJHSIG Adjectif Invariant en Genre débutant par un "h" aspiré
ADJIND Adjectif INDéfini
ADJINT Adjectif INTerrogatif
ADJINV Adjectif Invariant en Nombre et en Genre
ADJMIN Adjectif Masculin Invariant en Nombre
ADJMP Adjectif Masculin Pluriel
ADJMS Adjectif Masculin Singulier
ADJNUM Adjectif Numérique
ADJORD Adjectif Numérique Ordinal
ADJPIG Adjectif Pluriel Invariant en Genre
ADJSIG Adjectif Singulier Invariant en Genre
ADV Adverbe
COO Conjonction de Coordination
DETDEM Déterminant Démonstratif
DETDFS Déterminant Défini Féminin Singulier
DETDMS Déterminant Défini Masculin Singulier
DETDPIG Déterminant Défini Pluriel Invariant en Genre
DETIFS Déterminant Indéfini Féminin Singulier
DETIMS Déterminant Indéfini Masculin Singulier
DETPOSS Déterminant Possessif
INT Interjection
NCFIN Nom Commun Féminin Invariant en Nombre
NCFP Nom Commun Féminin Pluriel
NCFS Nom Commun Féminin Singulier
NCHFS Nom Commun Féminin Singulier débutant par un "h" aspiré
NCHSIG Nom Commun Singulier Invariant en Genre débutant par un "h" aspiré
NCHMIN Nom Commun Masculin Invariant en Nombre débutant par un "h" aspiré
NCHMS Nom Commun Masculin Singulier débutant par un "h" aspiré
NCI Nom Commun Invariant en Nombre et en Genre
NCMIN Nom Commun Masculin Invariant en Nombre
NCMP Nom Commun Masculin Pluriel
NCMS Nom Commun Masculin Singulier
NCPIG Nom Commun Pluriel Invariant en Genre
NCSIG Nom Commun Singulier Invariant en Genre
NHMIN Nom Masculin Invariant en Nombre débutant par un "h" aspiré
NPFIN Nom Propre Féminin Invariant en Nombre
NPFP Nom Propre Féminin Pluriel
NPFS Nom Propre Féminin Singulier
NPHFS Nom Propre Féminin Singulier débutant par un "h" aspiré
NPHMS Nom Propre Masculin Singulier débutant par un "h" aspiré
NPHSIG Nom Propre Singulier Invariant en Genre débutant par un "h" aspiré
NPI Nom Propre Invariant en Nombre et en Genre
NPMIN Nom Propre Masculin Invariant en Nombre
NPMP Nom Propre Masculin Pluriel
NPMS Nom Propre Masculin Singulier
NPPIG Nom Propre Pluriel Invariant en Genre

NPSIG Nom Propre Singulier Invariant en Genre
PCTFAIB Ponctuation Faible
PCTFORTE Ponctuation Forte
PDP Pronom Démonstratif Pluriel
PDS Pronom Démonstratif Singulier
PIFS Pronom Indéfini Féminin Singulier
PIFP Pronom Indéfini Féminin Pluriel
PII Pronom Indéfini Invariant en Genre et en Nombre
PIMP Pronom Indéfini Masculin Pluriel
PIMS Pronom Indéfini Masculin Singulier
PIPIG Pronom Indéfini Pluriel Invariant en Genre
PISIG Pronom Indéfini Singulier Invariant en Genre
PP Pronom Possessif
PPER1 P Pronom Personnel 1ère Personne du Pluriel
PPER1 S Pronom Personnel 1ère Personne du Singulier
PPER2 P Pronom Personnel 2ème Personne du Pluriel
PPER2 S Pronom Personnel 2ème Personne du Singulier
PPER3 P Pronom Personnel 3ème Personne du Pluriel
PPER3 S Pronom Personnel 3ème Personne du Singulier
PRFS Pronom Relatif Féminin Singulier
PRI Pronom Relatif Invariant en Genre et en Nombre
PRMS Pronom Relatif Masculin Singulier
PREP Préposition
SUB Conjonction de Subordination
SYMBOLE Symbole
UEUPH « l' » de style
VCONP1P Verbe Conditionnel Présent 1ère personne du Pluriel
VCONP2P Verbe Conditionnel Présent 2ème personne du Pluriel
VCONP3P Verbe Conditionnel Présent 3ème personne du Pluriel
VCONP1S Verbe Conditionnel Présent 1ère personne du Singulier
VCONP2S Verbe Conditionnel Présent 2ème personne du Singulier
VCONP3S Verbe Conditionnel Présent 3ème personne du Singulier
VIMPP1P Verbe Impératif Présent 1ère personne du Pluriel
VIMPP2P Verbe Impératif Présent 2ème personne du Pluriel
VIMPP2S Verbe Impératif Présent 2ème personne du Singulier
VINDF1P Verbe Indicatif Futur Simple 1ère personne du Pluriel
VINDF2P Verbe Indicatif Futur Simple 2ème personne du Pluriel
VINDF3P Verbe Indicatif Futur Simple 3ème personne du Pluriel
VINDF1S Verbe Indicatif Futur Simple 1ère personne du Singulier
VINDF2S Verbe Indicatif Futur Simple 2ème personne du Singulier
VINDF3S Verbe Indicatif Futur Simple 3ème personne du Singulier
VINDI1P Verbe Indicatif Imparfait 1ère Personne du Pluriel
VINDI2P Verbe Indicatif Imparfait 2ème personne du Pluriel
VINDI3P Verbe Indicatif Imparfait 3ème personne du Pluriel
VINDI1S Verbe Indicatif Imparfait 1ère personne du Singulier
VINDI2S Verbe Indicatif Imparfait 2ème personne du Singulier
VINDI3S Verbe Indicatif Imparfait 3ème personne du Singulier
VINDPS2P Verbe Indicatif Passé Simple 2ème personne du Pluriel
VINDPS3P Verbe Indicatif Passé Simple 3ème personne du Pluriel
VINDPS1S Verbe Indicatif Passé Simple 1ère personne du Singulier
VINDPS2S Verbe Indicatif Passé Simple 2ème personne du Singulier
VINDPS3S Verbe Indicatif Passé Simple 3ème personne du Singulier
VINDP1P Verbe Indicatif Présent 1ère personne du Pluriel
VINDP2P Verbe Indicatif Présent 2ème personne du Pluriel
VINDP3P Verbe Indicatif Présent 3ème personne du Pluriel
VINDP1S Verbe Indicatif Présent 1ère personne du Singulier

VINDP2S Verbe Indicatif Présent 2ème personne du Singulier
VINDP3S Verbe Indicatif Présent 3ème personne du Singulier
VINF Verbe à l' Infinitif
VPARPFP Verbe Participe Passé Féminin Pluriel
VPARPFS Verbe Participe Passé Féminin Singulier
VPARPMS Verbe Participe Passé Masculin Singulier
VPARPMP Verbe Participe Passé Masculin Pluriel
VPARPRES Verbe Participe Présent
VSUBI3S Verbe Subjonctif Imparfait 3ème personne du Singulier
VSUBP1P Verbe Subjonctif Présent 1ère personne du Pluriel
VSUBP2P Verbe Subjonctif Présent 2ème personne du Pluriel
VSUBP3P Verbe Subjonctif Présent 3ème personne du Pluriel
VSUBP1S Verbe Subjonctif Présent 1ère personne du Singulier
VSUBP2S Verbe Subjonctif Présent 2ème personne du Singulier
VSUBP3S Verbe Subjonctif Présent 3ème personne du Singulier

Annexe 2 : Extrait d'un fichier étiqueté par Cordial Analyseur

en effet	en effet	ADV
on	on	PPER3S
peut	pouvoir	VINDP3S
commencer	commencer	VINF
bon	bon	ADJMS
alors	alors	ADV
d'abord	d'abord	ADV
madame	madame	NCFIN
je	je	PPER1S
vais	aller	VINDP1S
vous	vous	PPER2P
vous	vous	PPER2P
poser	poser	VINF
quelques	quelque	ADJIND
petites	petit	ADJFP
questions	question	NCFP
préliminaires	préliminaire	ADJPIG
euh	euh	INT
depuis	depuis	PREP
combien de	combien de	ADJIND
temps	temps	NCMIN
habitez	habiter	VINDP2P
-vous	vous	PPER2P
Orléans	Orléans	NPHSIG
?	?	PCTFORTE
dix	dix	ADJNUM
ans	an	NCMP
depuis	depuis	PREP
dix	dix	ADJNUM
ans	an	NCMP
oui	oui	ADV
et	et	COO
qu'	que	PRI
est	être	VINDP3S
ce	ce	PDS
qui	qui	PRI
vous	vous	PPER2P
a	avoir	VINDP3S
amené	amener	VINDP2P
à	à	PREP
vivre	vivre	VINF
ici	ici	ADV
?	?	PCTFORTE
la	le	DETDFS
situation	situation	NCFS
de	de	PREP
mon	mon	DETPOSS
fils	fils	NCMIN

Annexe 3 : Etiquettes utilisées dans les premières expériences

	étiquettes	nombre d'étiquettes selon les niveaux			Modifications par rapport à Cordial
		L0	L1	L2	
nom	NCMS, NCMP, NCFS, NCFP, NCI, NP.	2	5	5	suppression du « h » aspiré, les noms invariants sont rassemblés sous une étiquette, les noms propres sont réunis en une seule étiquette
adjectif	ADJMS, ADJMP, ADJFS, ADJFP, ADJI	1	5	5	suppression du « h » aspiré, les adjectifs invariants sont réunis en une seule étiquette, les adjectifs ordinaux sont mis parmi les adjectifs, les adjectifs cardinaux sont dans la nouvelle étiquette CH
verbe	V1+2+3S+PINDP+I+F+PS, V1+2+3S+SUB, V1+2+3S+CON, VINF, VPARPRES, VM+FS+PARP	1	45	45	Les temps pour le subjonctif, le conditionnel et l'impératif ne sont plus précisés, tout est regroupé en une seule étiquette pour chacun de ces modes.
adverbe	ADV	1	1	1	
pronom	PM+FS+P+IREL+INT+PER+POSS+DEM+IND	1	5	29	les démonstratifs, possessifs et relatifs ont été détaillés en genre et nombre ; les pronoms indéfinis invariants sont regroupés en une seule étiquette au lieu de trois
déterminant	DETMS+FS+P+IDEM+INT+POSS+DEM+IND+DEF	1	4	20	les démonstratifs, indéfinis et possessifs ont été détaillés en genre et nombre pour conserver une cohérence vis à vis des autres catégories de déterminants; la catégorie des interrogatifs a été ajoutée
chiffre	CH	1	1	1	
préposition	PREP	1	1	1	
conjonction	CONJSUB, CONJCOO	1	1	2	
interjection	INT	1	1	1	
mot inconnu	MI	1	1	1	
présentateur	PRES	1	1	1	
ponctuation	PCT	1	1	1	Les deux étiquettes de Cordial Analyseur sont réduites en une seule
total		14	72	114	

Annexe 4 : Extrait du corpus de référence utilisé au cours des premières expériences

oui	oui	ADV	ADV	ADV
en_effet	en_effet	ADV	ADV	ADV
on	on	P	P3I	P3IPER
peut	pouvoir	V	V3SINDP	V3SINDP
commencer	commencer	V	VINF	VINF
bon	bon	INT	INT	INT
alors	alors	ADV	ADV	ADV
d'abord	d'abord	ADV	ADV	ADV
madame	madame	N	NCFS	NCFS
je	je	P	P1S	P1SPER
vais	aller	V	V1SINDP	V1SINDP
vous	vous	P	P2P	P2PPER
vous	vous	P	P2P	P2PPER
poser	poser	V	VINF	VINF
quelques	quelque	DET	DETFP	DETFPIND
petites	petit	ADJ	ADJFP	ADJFP
questions	question	N	NCFP	NCFP
préliminaires	préliminaire	ADJ	ADJFP	ADJFP
euh	euh	INT	INT	INT
depuis	depuis	PREP	PREP	PREP
combien_de	combien_de	ADV	ADJV	ADJV
temps	temps	N	NCI	NCI
habitez	habiter	V	V2PINDP	V2PINDP
vous	vous	P	P2P	P2PPER
Orléans	Orléans	N	NP	NP
dix	dix	CH	CH	CH
ans	an	N	NCMP	NCMP
depuis	depuis	PREP	PREP	PREP
dix	dix	CH	CH	CH
ans	an	N	NCMP	NCMP
oui	oui	ADV	ADV	ADV
et	et	CONJ	CONJ	CONJCOO
qu'	que	P	PI	PIINT
est	être	V	V3SINDP	V3SINDP
ce	ce	P	PMS	PMSDEM
qui	qui	P	PI	PIREL
vous	vous	P	P2P	P2PPER
a	avoir	V	V3SINDP	V3SINDP
amené	amener	V	VMSPARP	VMSPARP
à	à	PREP	PREP	PREP
vivre	vivre	V	VINF	VINF
ici	ici	ADV	ADV	ADV
la	le	DET	DETFS	DETFSDEF
situation	situation N	NCFS	NCFS	
de	de	PREP	PREP	PREP
mon	mon	DET	DETMS	DETMSPOSS
fils	fils	N	NCI	NCI

Annexe 5 : Exemple d'étiquetage effectué par Cordial et adapté pour notre corpus

mot du corpus	lemmes	étiquette Cordial	étiquette à nous			sens
			L0	L1	L2	
oui	Oui	ADV	ADV	ADV	ADV	adverbe
en	en	PREP	(en_effet) ADV	ADV	ADV	adverbe
effet	effet	NCMIN				
on	on	PPER3S	P	P3I	P3IPER	pronom 3 pers invariable personnel
peut	pouvoir	VINDP3S	V	V3SINDP	V3SINDP	verbe 3 pers. sing. Indicatif présent
commencer	commencer	VINF	V	VINF	VINF	verbe à l'infinitif
bon	bon	ADJMS	INT	INT	INT	interjection
alors	alors	ADV	ADV	ADV	ADV	adverbe
d'abord	d'abord	ADV	ADV	ADV	ADV	adverbe
madame	madame	NCFIN	N	NCFS	NCFS	nom commun féminin singulier
je	je	PPER1S	P	P1S	P1SPER	pronom 1 pers.sing.personnel
vais	aller	VINDP1S	V	V1SINDP	V1SINDP	verbe 1 pers. sing. indicatif présent
vous	vous	PPER2P	P	P2P	P2PPER	pronom 2 pers. pluriel personnel
vous	vous	PPER2P	P	P2P	P2PPER	pronom 2 pers. pluriel personnel
poser	poser	VINF	V	VINF	VINF	verbe à l'infinitif
quelques	quelque	ADJIND	DET	DETFP	DETFPIND	pronom féminin pluriel indéfini
petites	petit	ADJFP	ADJ	ADJFP	ADJFP	adjectif féminin pluriel
questions	question	ADJFP	N	NCFP	NCFP	nom commun féminin pluriel
préliminaires	préliminaire	ADJPIG	ADJ	ADJFP	ADJFP	adjectif féminin pluriel
euh	euh	INT	INT	INT	INT	interjection
depuis	depuis	PREP	PREP	PREP	PREP	préposition
combien de	combien de	ADJIND	ADV	ADV	ADV	adverbe
temps	temps	NCMIN	N	NCI	NCI	nom commun invariable
habitez	habiter	VINDP2P	V	V2PINDP	V2PINDP	verbe 2 pers. pluriel indicatif présent
vous	vous	PPER2P	P	P2P	P2PPER	pronom 2 pers. pluriel personnel
Orléans	Orléans	NPHSIG	N	NP	NP	nom propre

?	?	PCTFORTE	PCT	PCT	PCT	
dix	dix	ADJNUM	CH	CH	CH	chiffre
ans	an	NCMP	N	NCMP	NCMP	nom commun masculin pluriel
depuis	depuis	PREP	PREP	PREP	PREP	préposition
dix	dix	ADJNUM	CH	CH	CH	chiffre
ans	an	NCMP	N	NCMP	NCMP	nom commun masculin pluriel
oui	oui	ADV	ADV	ADV	ADV	adverbe
et	et	COO	CONJ	CONJ	CONJCOO	conjonction de coordination
qu'	que	PRI	P	PI	PIINT	pronom invariable interrogatif
est	être	VINDP3S	V	V3SINDP	V3SINDP	verbe 3 pers singulier indicatif présent
ce	ce	PDS	P	PMS	PMSDEM	pronom masculin singulier démonstratif
qui	qui	PRI	P	PI	PIREL	pronom invariable relatif
vous	vous	PPER2P	P	P2P	P2PPER	pronom 2 pers. pluriel personnel
a	avoir	VINDP3S	V	V3SINDP	V3SINDP	verbe 3 pers singulier indicatif présent
amené	amener	VINDP2P	V	VMSPARP	VMSPARP	verbe masculin singulier participe passé
à	à	PREP	PREP	PREP	PREP	préposition
vivre	vivre	VINF	V	VINF	VINF	verbe à l'infinitif
ici	ici	ADV	ADV	ADV	ADV	adverbe
?	?	PCTFORTE	PCT	PCT	PCT	ponctuation
la	le	DETDFS	DET	DETFS	DETFSDEF	déterminant féminin singulier défini
situation	situation	NCFS	N	NCFS	NCFS	nom commun féminin singulier
de	de	PREP	PREP	PREP	PREP	préposition
mon	mon	DETPOSS	DET	DETMS	DETMSPOSS	déterminant masculin singulier possessif
fils	fils	NCMIN	N	NCI	NCI	nom commun invariable

Annexe 6 : Jeu d'étiquettes morpho-syntaxiques d'ESLO

	ESLO			
POS	+MORPH	+SOUS-TYPE	NOMBRE D'ETIQUETTES	
			POS	COMPLETES
N	NCMS/FS/MP/FP	NCMS/FS/MP/FP	1	4
	NP	NP		1
ADJ	ADJMS/FS/MP/FP	ADJMS/FS/MP/FP	1	4
DET	DETMS/FS/P	DETMS/FS/PDEM	1	3
	DETMS/FS/P1-3S	DETMS/FS/P1-3SPOSS		9
	DETS/P1-3P	DETS/P1-3PPOSS		6
	DETMS/FS/P	DETMS/FS/PDEF		3
	DETMS/FS/P/I	DETMS/FS/P/IIND		4
	DETMS/FS/MP/FP	DETMS/FS/MP/FPINT		4
P	PMS/FS/MP/FP/I	PMS/FS/MP/FP/IREL	1	5
	PMS/FS/MP/FP/I	PMS/FS/MP/FP/IINT		5
	PMS/FS/MP/FP/I	PMS/FS/MP/FP/IDEM		5
	P1-3S/MS/FS/P/MP/FP	P1-3S/MS/FS/P/MP/FPPERSUJ		8
	P1-3S/MS/FS/P	P1-3S/MS/FS/PPERCOMPL		9
	P1-3S/MS/FS/P/MP/FP	P1-3S/MS/FS/P/MP/FPPERTON		9
	P3S	P3SPERADV		1
	PMS/FS/MP/FP1-3S	PMS/FS/MP/FP1-3SPOSS		12
	PMS/FS/P1-3P	PMS/FS/P1-3PPOSS		9
	PMS/FS/MP/FP	PMS/FS/MP/FPIND		4
V	V1-3S/PINDP/F/I/PS/CON/PAUX	V1-3S/PINDP/F/I/PS/CON/PAUX	1	36
	VMS/FS/MP/FPPP	VMS/FS/MP/FPPP		4
	VMS/FS/MP/FPPP	VMS/FS/MP/FPPPPAS		4
	V1-3S/PSUB	V1-3S/PSUB		6
	V1-2S/PIMP	V1-2S/PIMP		3
	VINF	VINF		1
	VPPRES Verbe au Participe Présent	VPPRES Verbe au Participe Présent		1
ADV	ADV	ADV/ADVNEG	1	2
PREP	PREP	PREP	1	1
CONJ	CONJ	CONJCOO/SUB	1	2
INT	INT	INT	1	1
PRES	PRES	PRES	1	1
MD	MD	MD/MDEUH/MDINT	1	3
MI	MI	MI	1	1
CH	CH	CH	1	1
UEUPH	UEUPH	UEUPH	1	1
PCT	PCT	PCT	1	1
			16	173

Annexe 7 : Tableau comparatif de jeux d'étiquettes de quatre étiqueteurs

Dister	TCOF	DisMo	ESLO
N+m/f/s/p	NOM (nom commun), NOM :sig (sigle), NOM :trc (nom commun tronqué)	NOM:com	NC+M/F/S/P
NPr	NAM (nom propre), NAM :sig (sigle), NAM :trc (nom propre tronqué)	NOM:prop, NOM:acr	NP
A+m/f/s/p	ADJ, ADJ :trc (adjectif tronqué)	ADJ:adj	ADJ+M/F/S/P
DET+Ddem+m/f/s/p	DET:dem	DET:dem	DET+M/F/S/P+DEM
DET+Dpos1-3s/1-3p+m/f/s/p	DET:pos	DET:pos	DET+M/F/S/P1-3S/P+POSS
DET+Ddef+m/f/s/p, DET+Ddef+Prepdet+m/f/s/p	DET:def	DET:def	DET+M/F/S/P+DEF
DET+Dind+m/f/s/p	DET:ind	DET:ind	DET+M/F/S/P+IND
DET+Dexi (déterminans exclamatif et interrogatif)	DET:int		DET+M/F/S/P+INT
	DET :pre (pré-déterminant tout (le))		
	DET :par (déterminant partitif du)		
DET+Dnum		DET:num	
PRO+Prel+Preppro+m/f/s/p	PRO:rel	PRO:rel	P+M/F/S/P/I+REL
	PRO:int		P+M/F/S/P/I+INT
PRO+Pdem+m/f/s/p	PRO:dem	PRO:dem	P+M/F/S/P/I+DEM
PRO+Ppers+m/f/s/p	PRO:cls	PRO:sjt (pronom pers. sujet)	P+1-3M/F/S/P+PERSUJ
	PRO :clsi (clitique sujet impersonnel)		
	PRO:clo	PRO:nprp (pronom clitique non prépositionnel direct)	P+1-3M/F/S/P+PERCOMPL
	PRO:ton	PRO:ton	P+1-3M/F/S/P+PERTON
PRO+Ppos1-3s/1-3p+m/f/s/p	PRO:poss	PRO:pos	P+M/F/S/P1-3S/P+POSS
PRP+Pind+Preppro+m/f/s/p	PRO:ind	PRO:ind	P+M/F/S/P+IND
		PRO:prp (pronom clitique prépositionnel indirect)	P3SPERADV
V+1-3s/1-3p+B/C/F/G/I/J/K/P/S/T/Y/W	VER:pre, VER:futu, VER:impf, VER:simp, VER:cond	VER:cond, VER:futu, VER:impf, VER:pres, VER:simp (passé simple),	V+1-3S/PINDP/F/I/PS/CON

Dister	TCOF	DisMo	ESLO
	VER :pper (verbe au participe passé), VER :ppre (verbe au participe présent)	VER:ppa, VER:ppe (verbe participe passé et présent)	V+M/F/S/P/PP, V+PPRES
	VER :subi (verbe au subjonctif imparfait), VER :subp (verbe au subjonctif présent)	VER:subi, VER:subp (verbe subjonctif imparfait et présent)	V+1-3S/PSUB
	VER:impe	VER:impe	V+1-2S/PIMP
	VER:infi	VER:inf	V+INF
	AUX:cond, AUX:futu, AUX:impe, AUX:impf, AUX:infi, AUX:pper, AUX:ppre, AUX:pres, AUX:simp, AUX:subi, AUX:sbbp	VER:xxx:aux	V+1-3S/PINDPAUX
	VER (verbe sans flexion voilà)		
	VER :trc verbe tronqué		
		VER:xxx:pred (je suis gentil)	
ADV	ADV	ADV:adv, ADV:comp, ADV:deg, ADV:int, ADV:neg	ADV, ADV+NEG
PREP+eff, PREP+Preppro+m/f/s/p	PRP	PRP	PREP
PREP+Prepdet+m/f/s/p	PRP :det (préposition/déterminant) => article contracté	PRP:det	
CONJC, CONJS	KON	CON:coo, CON:sub, CONN	CONJ+COO; CONJ+SUB
INTJ	INT (interjection et particules discursives)	INTJ, MD	MD, MD+EUH, MD+INT
	TRC (amorces de mots)	AMO	MI
	NUM (numéral)	NUM:num	CH
	EPE (épenthétique)		UEUPH
	PRT :int (particule interrogative est ce que)		
	ETR (mots étrangers)		
	FNO forme noyau (oui, non, d'accord etc.)	PARA (discours para-verbal)	
	LOC (locuteur)		
	MLT (multi-transcription)		
	SYM (symbole)		PCT
PFX		PFX préfixe	

Dister	TCOF	DisMo	ESLO
		CORR-B (autocorrection initiale erronée), CORR-I (autocorrection interne corrigé)	
		REP-B, REP-I	
		SIL:b (pause brève), SIL:l (pause longue), SIL:s (pause avec prise de souffle)	
		HESI (hésitation) euh	
			PRES
167	62	56	172

Annexe 8 : Extrait du corpus ESLO annoté manuellement en chunks FTB adaptés à l'oral

	POS	CHUNK	
en_effet	ADV	B-AdP	
on	CLS	B-NP	
peut	V	B-VN	
commencer	VINF	B-VN	
bon	I	B-AdP	
alors	I	B-AdP	
d'abord	ADV	B-AdP	
madame	NC	B-NP	
je	CLS	B-NP	
vais	V	B-VN	
vous	CLO	B-NP	
v-	UKNOWN		B-UKNOWN
vous	CLO	B-NP	
poser	VINF	B-VN	
quelques	DET	B-NP	
petites	ADJ	I-NP	
questions	NC	I-NP	
préliminaires	ADJ	I-NP	
euh	I	B-AdP	
depuis	P	B-PP	
combien	PROWH	I-PP	
de	P	B-PP	
temps	NC	I-PP	
habitez	V	B-VN	
-vous	CLS	B-NP	
Orléans	NPP	B-NP	
dix	DET	B-NP	
ans	NC	I-NP	
depuis	P	B-PP	
dix	DET	I-PP	
ans	NC	I-PP	
oui	I	B-AdP	
et	CC	B-CONJ	
qu'	PROREL	B-NP	
est	V	B-VN	
ce	PRO	B-NP	
qui	PROREL	B-NP	
vous	CLO	B-NP	
a	V	B-VN	
amené	VPP	I-VN	
à	P	B-PP	
vivre	VINF	I-PP	
ici	ADV	B-AdP	
la	DET	B-NP	
situation	NC	I-NP	

de	P	B-PP
mon	DET	I-PP
fils	NC	I-PP
ah_oui	I	B-AdP
et	CC	B-CONJ
vous	CLS	B-NP
vous	CLR	B-NP
plaisez	V	B-VN
à	P	B-PP
Orléans	NPP	I-PP
oh	I	B-AdP
je	CLS	B-NP
me	CLR	B-NP
plais	V	B-VN
bien	ADV	B-AdP
oui	I	B-AdP
oui	I	B-AdP
oui	I	B-AdP
pourquoi	ADVWH	B-AdP
vous	CLS	B-NP
dites	V	B-VN
cela	PRO	B-NP

Annexe 9 : Extrait du corpus ESLO annoté manuellement en chunks adaptés à l'oral

	POS(non corrigé)	CHUNK(manuel)	CHUNK(SEM appris)
oui	ADV	B-IntP	B-AdP
en_effet	ADV	B-AdP	B-AdP
on	CLS	B-NP	B-NP
peut	V	B-VN	B-VN
commencer	VINF	B-VN	B-VN
bon	ADJ	B-IntP	B-AP
alors	ADV	B-IntP	B-AdP
d'abord	NC	B-AdP	B-NP
madame	V	B-NP	B-VN
je	CLS	B-NP	B-NP
vais	V	B-VN	B-VN
vous	CLS	B-NP	B-NP
v-	V	B-UNKNOWN	B-VN
vous	CLO	B-NP	B-NP
poser	VINF	B-VN	B-VN
quelques	DET	B-NP	B-NP
petites	ADJ	I-NP	I-NP
questions	NC	I-NP	I-NP
préliminaires	ADJ	I-NP	I-NP
euh	NC	B-IntP	B-IntP
depuis	P	B-PP	B-PP
combien	NC	I-PP	I-PP
de	P	B-PP	B-PP
temps	NC	I-PP	I-PP
habitez	V	B-VN	B-VN
-vous	ADJ	B-NP	B-AP
Orléans	NPP	B-NP	B-NP
dix	DET	B-NP	B-NP
ans	NC	I-NP	I-NP
depuis	P	B-PP	B-PP
dix	DET	I-PP	I-PP
ans	NC	I-PP	I-PP
oui	NC	B-IntP	B-IntP
et	CC	B-CONJ	B-CONJ
qu'	PROREL	B-NP	B-NP
est	V	B-VN	B-VN
ce	PRO	B-NP	B-NP
qui	PROREL	B-NP	B-NP
vous	CLO	B-NP	B-NP
a	V	B-VN	B-VN
amené	VPP	I-VN	I-VN
à	P	B-PP	B-PP
vivre	VINF	I-PP	I-PP
ici	ADV	B-AdP	B-AdP
la	DET	B-NP	B-NP

	POS(non corrigé)	CHUNK(manuel)	CHUNK(SEM appris)
situation	NC	I-NP	I-NP
de	P	B-PP	B-PP
mon	DET	I-PP	I-PP
fils	NC	I-PP	I-PP
ah_oui	NC	B-IntP	B-IntP
et	CC	B-CONJ	B-CONJ
vous	CLS	B-NP	B-NP
vous	CLO	B-NP	B-NP
plaisez	V	B-VN	B-VN
à	P	B-PP	B-PP
Orléans	NPP	I-PP	I-PP
oh	V	B-IntP	B-IntP
je	CLS	B-NP	B-NP
me	CLR	B-NP	B-NP
plais	V	B-VN	B-VN
bien	ADV	B-AdP	B-AdP
oui	VPP	B-IntP	B-IntP
oui	ADV	B-IntP	B-IntP
oui	ADV	B-IntP	B-IntP
pourquoi	ADVWH	B-AdP	B-AdP
vous	CLS	B-NP	B-NP
dites	VPP	B-VN	B-VN
cela	PRO	B-NP	B-NP

Annexe 10 : Typologie des éléments annotés par la cascade CasDen

L'information concernant le locuteur : *pers*	*pers.speaker*	la personne interviewée
	pers.spouse	son époux ou épouse
	pers.child	ses enfants
	pers.parent	ses autres liens de parenté
Identité : *identity*	*identity.name*	le nom
	identity.addr	l'adresse
	identity.wedding	le mariage
	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
L'information concernant le travail : *work*	*work.occupation*	le métier
	work.field	le domaine professionnel
	work.location	le lieu de travail
	work.business	l'entreprise
L'information concernant les engagements de la personne dans la vie sociale : *involvement*	*involvement.military*	l'engagement militaire
	involvement.voluntary	l'engagement associatif
	involvement.school	l'engagement scolaire
	involvement.tradeunion	l'appartenance syndicale
Les voyages effectués : *trip*	*trip.study*	le voyage d'études
	trip.holiday	le voyage de vacances
	trip.work	le voyage professionnel
L'information concernant les études : *study*	*study.location*	le lieu d'études
	study.degree	le diplôme
	study.edu	l'établissement d'études

Annexe 11 : Extrait de l'entretien 273 du corpus ESLO1 annoté en indices d'identification par la cascade CasDen

<Episode>
<Section type="report" startTime="0" endTime="3840.145">
<Turn startTime="0" endTime="25.237" speaker="spk1">
<Sync time="0"/>
<Sync time="0.429"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
parce que
<Sync time="1.932"/>
il ne pourra pas apporter des vipères
<Sync time="3.768"/>
<Sync time="4.202"/>
c'est un lieu ou euh
<Sync time="5.394"/>
le premier lieu de camps
<Sync time="6.565"/>
<Sync time="7.231"/>
qui a été avait été décidé
<Sync time="9.0"/>
<Sync time="9.454"/>
notre enfin le seul et unique chef qui reste puisque l'autre s'est débiné le pauvre
<Sync time="13.417"/>
<Sync time="14.157"/>
à <DE type="identity.name"><ENT type="pers.hum">Hervé</ENT></DE> dit ça c'est un salaud
<Sync time="15.64"/>
<Sync time="16.308"/>
ça maman moi si <ENT type="time.date.rel">un jour</ENT> <DE type="pers.speaker">je suis<DE type="work.occupation"> chef</DE></DE> je crois que jamais je ferai ça
<Sync time="19.003"/>
<Sync time="19.652"/>
ah j'ai dis ne dis pas fontaine je ne boirais pas de ton eau
<Sync time="22.566"/>
<Sync time="23.329"/>
ne mais oh c'est vrai <DE type="identity.name"><ENT type="pers.hum">Philippe</ENT></DE> a été
</Turn>
<Turn speaker="spk1 spk2" startTime="25.237" endTime="27.598">
<Sync time="25.237"/>
<Who nb="1"/>
m'enfin c'est moche quand même hein
<Who nb="2"/>
où est ce qu'ils partent au
<Event desc="pi" type="pronounce" extent="instantaneous"/>
</Turn>
<Turn speaker="spk1" startTime="27.598" endTime="59.267">
<Sync time="27.598"/>

alors ils partaient à la<ENT type="loc.admi"> Roche Blanche</ENT> à côté de<ENT type="loc.admi"> Clermont-Ferrand</ENT>
<Sync time="30.488"/>
<Sync time="30.989"/>
et puis maintenant ils vont je me rappelle même pas l'adresse on nous a donné les petits papiers <ENT type="time.date.rel">hier soir</ENT> dans nos boîtes aux lettres
<Sync time="36.641"/>
à <ENT type="amount.phy.len">cinq kilomètres</ENT>
<Sync time="38.172"/>
parce qu'en arrivant sur le lieu de camps <ENT type="time.date.rel">dimanche</ENT>
<Sync time="40.39"/>
ils ont trouvé enfin euh je peut être que j'exagère mais énormément de vipères
<Sync time="45.26"/>
<Sync time="45.718"/>
mais mon époux leurs avait dit à la réunion à la maison mon époux qui connait parfaitement l'<ENT type="loc.admi">Auvergne</ENT>
<Sync time="51.015"/>
<Sync time="51.497"/>
avait dit si j'ai
<Sync time="52.885"/>
peux me permettre de donner un petit entre filet il avait dit à <DE type="identity.name"><ENT type="pers.hum">Philippe</ENT></DE>
<Sync time="55.823"/>
<Sync time="56.4"/>
surtout méfiez-vous des vipères
<Sync time="58.165"/>
la preuve c'est que
</Turn>
<Turn speaker="spk2" startTime="59.267" endTime="61.227">
<Sync time="59.267"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
ils allaient en recevoir
<Event desc="pi" type="pronounce" extent="instantaneous"/>
</Turn>
<Turn speaker="spk1" startTime="61.227" endTime="80.19">
<Sync time="61.227"/>
oh oui parce que moi <DE type="identity.name"><ENT type="pers.hum">Hervé</ENT></DE> du coup dit tu sais je suis un peu refroidit <DE type="identity.name"><ENT type="pers.hum">Hervé</ENT></DE> à horreur de ça
<Sync time="65.877"/>
alors <ENT type="time.hour">tout à l'heure</ENT> je devais aller avec lui acheter des bâtons de glace
<Sync time="68.763"/>
<Sync time="69.431"/>
j'ai dis écoute mon petit chou moi je peux pas y arriver ma pauvre femme de ménage doit être fatiguée
<Sync time="73.103"/>
enfin en <ENT type="time.hour">trois heures et demie</ENT> <ENT type="time.date.abs">ce matin</ENT> elle a pas fait le travail représentant <ENT type="time.hour">une heure</ENT>
<Sync time="77.233"/>
quand j'ai vu <ENT type="time.hour">tout à l'heure</ENT>
<Event desc="pi" type="pronounce" extent="instantaneous"/>

<Sync time="79.093"/>
oh j'ai eu mon coup de dépression
</Turn>
<Turn speaker="spk2" startTime="80.19" endTime="81.526">
<Sync time="80.19"/>
et puis moi je viens vous déranger
</Turn>
<Turn speaker="spk1" startTime="81.526" endTime="91.69">
<Sync time="81.526"/>
mais non pas du tout non non j'ai fini vous savez j'ai une presse alors je ne m'ennuie pas d'affolement
<Sync time="86.014"/>
j'ai dis<DE type="identity.name"><ENT type="pers.hum"> Hervé</ENT></DE> je vais venir faire ça avec moi mais ne me saoule pas
<Sync time="89.09"/>
<Sync time="89.758"/>
il est d'une exitation écoutez <ENT type="pers.hum.tit">madame</ENT>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
</Turn>
<Turn speaker="spk2" startTime="91.69" endTime="92.959">
<Sync time="91.69"/>
alors ils partent <ENT type="time.date.abs">ce soir</ENT>
</Turn>
<Turn speaker="spk1" startTime="92.959" endTime="96.112">
<Sync time="92.959"/>
oh il dit quelle chance quel débarras quand tu vas plus m'avoir j'ai dis oui ah
</Turn>
<Turn speaker="spk2" startTime="96.112" endTime="96.856">
<Sync time="96.112"/>
<Event desc="rire" type="noise" extent="instantaneous"/>
</Turn>
<Turn speaker="spk1" startTime="96.856" endTime="98.478">
<Sync time="96.856"/>
ah ça j'ai dis oui ça vraiment
</Turn>
<Turn speaker="spk2" startTime="98.478" endTime="102.585">
<Sync time="98.478"/>
si vous saviez que moi j'en ai une au camps de <DE type="identity.name"><ENT type="pers.hum">Guy</ENT></DE> que j'ai reçu une lettre de la cheftaine <ENT type="time.date.abs">ce matin</ENT> que ça va pas du tout
</Turn>
<Turn speaker="spk1" startTime="102.585" endTime="103.234">
<Sync time="102.585"/>
laquelle ?
</Turn>
<Turn speaker="spk2" startTime="103.234" endTime="103.787">
<Sync time="103.234"/>
ben <DE type="identity.name"><ENT type="pers.hum">Isabelle</ENT></DE>
</Turn>
<Turn speaker="spk1" startTime="103.787" endTime="105.509">

<Sync time="103.787"/>
<DE type="identity.name"><ENT type="pers.hum">Isabelle</ENT></DE>
<Sync time="104.479"/>
<Sync time="105.032"/>
comment ça se fait ?
</Turn>
<Turn speaker="spk2" startTime="105.509" endTime="107.012">
<Sync time="105.509"/>
il ne s'entend pas avec sa cheftaine ça c'est
</Turn>
<Turn speaker="spk1" startTime="107.012" endTime="107.828">
<Sync time="107.012"/>
aïe aïe aïe
</Turn>
<Turn speaker="spk2" startTime="107.828" endTime="111.434">
<Sync time="107.828"/>
et puis alors elle est parti en disant en pleurant moi la cheftaine je lui en ai parlé <ENT type="time.date.abs">le matin</ENT> avant de partir
</Turn>
<Turn speaker="spk2 spk1" startTime="111.434" endTime="114.324">
<Sync time="111.434"/>
<Who nb="1"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
alors ça va pas du tout
<Who nb="2"/>
oh ça c'est moche
</Turn>
<Turn speaker="spk2" startTime="114.324" endTime="124.451">
<Sync time="114.324"/>
alors j'ai appelé <ENT type="time.date.rel">à midi</ENT> au téléphone
<Sync time="116.256"/>
et puis ils sont partis en excursion à<ENT type="loc.admi"> Bourges</ENT>
<Sync time="118.479"/>
mon mari me dit bon puisque c'est ça il dit la cheftaine elle commence à nous faire suer euh tu prends la voiture il dit tu va les chercher
<Sync time="123.015"/>
eh ben il me dit tu va les chercher
</Turn>
<Turn speaker="spk2 spk3" startTime="124.451" endTime="129.178">
<Sync time="124.451"/>
<Who nb="1"/>
oh mais c'est
<Event desc="pi" type="pronounce" extent="instantaneous"/>
on a réglé avec
<Event desc="pi" type="pronounce" extent="instantaneous"/>
de la
<Who nb="2"/>
alors sur ce je dis bon ben écoute
</Turn>

<Turn speaker="spk2" startTime="129.178" endTime="129.827">
<Sync time="129.178"/>
de de la
<Event desc="pi" type="pronounce" extent="instantaneous"/>
</Turn>
<Turn speaker="spk1" startTime="129.827" endTime="137.568">
<Sync time="129.827"/>
vous savez elles sont pas
<Sync time="130.595"/>
les les encadrements sont pas faciles
<Sync time="132.293"/>
vous savez je ramène les miens du camps louveteau
<Sync time="134.372"/>
eh bien <ENT type="pers.hum.tit">madame</ENT> je peux pas me permettre de juger parce que je trouve que c'est déjà du
</Turn>
<Turn speaker="spk1 spk2" startTime="137.568" endTime="140.31">
<Sync time="137.568"/>
<Who nb="1"/>
dévouement je leurs trouverai d'emmener des garçons
<Who nb="2"/>
oui d'accord
</Turn>
<Turn speaker="spk1" startTime="140.31" endTime="191.131">
<Sync time="140.31"/>
mais il faut quand même
<Sync time="141.579"/>
j'en di- je le disais avec <ENT type="pers.hum">le Père Besançon</ENT> qu'on a vu <ENT type="time.date.rel">cette semaine</ENT> j'ai dis<ENT type="pers.hum"> Père</ENT>
<Sync time="144.97"/>
quand on accepte de prendre la responsabilité d'enfants
<Sync time="148.118"/>
c'est quand même très grâve encore plus <ENT type="time.hour">à l'heure actuelle</ENT> qu'autrement hein
<Sync time="152.201"/>
ben écoutez vraiment moi j'ai été
<Sync time="153.97"/>
mon époux est lui pourtant ce sont des petites choses auquelles les femmes prêtent plus d'importance si vous voulez
<Sync time="159.718"/>
mais vraiment quand vous voyez le vocabulaire des chansons de corps de garde
<Sync time="164.206"/>
que m'a raconté<DE type="identity.name"><ENT type="pers.hum"> Thierry</ENT></DE> <ENT type="time.date.abs">dimanche matin</ENT> dans le plus grand secret
<Sync time="167.521"/>
parce que c'était tellement grossier qu'il a
<Sync time="169.744"/>
sans même comprendre tout ce qu'il nous a raconté
<Sync time="172.306"/>
écoutez j'ai dis vraiment il y en a un qui dans la bande

<Sync time="175.377"/>
dans tous les camps louveteau dans toutes les collectivités
<Sync time="178.291"/>
ça se produit mais c'est ce que je disais au <ENT type="pers.hum">Père Besançon</ENT>
<Sync time="180.724"/>
les cheftaines ont quand même une oreille attentive à avoir quand une fois on a chanté cette grossièreté là elle aurait dû faire la ronde
<Sync time="187.525"/>
ça continuait tout le camp ah ben mon <DE type="identity.name"><ENT type="pers.hum">Thierry</ENT></DE> ah ben
<Sync time="190.268"/>
</Turn>
<Turn speaker="spk2" startTime="191.131" endTime="192.538">
<Sync time="191.131"/>
ah oui ça moi je me souviens toujours
</Turn>
<Turn speaker="spk1 spk2" startTime="192.538" endTime="198.242">
<Sync time="192.538"/>
<Who nb="1"/>
il a<ENT type="amount.phy.age"> huit ans</ENT>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
<Who nb="2"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
si vous saviez toutes les histoires qu'il racontait c'était pas croyable
</Turn>
<Turn speaker="spk2" startTime="198.242" endTime="200.083">
<Sync time="198.242"/>
vous savez au camp de pionniers <ENT type="time.date.rel">cette année</ENT> <ENT type="time.date.rel">cette année</ENT>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
</Turn>
<Turn speaker="spk3" startTime="200.083" endTime="201.132">
<Sync time="200.083"/>
j'avais dis au Père
</Turn>
<Turn speaker="spk2" startTime="201.132" endTime="209.627">
<Sync time="201.132"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
ils ont eu des ennuis
<Event desc="pi" type="pronounce" extent="instantaneous"/>
<Sync time="203.875"/>
et euh vous savez ils se entre garçons dans les camps de rangers et tout ça ben
<Sync time="207.862"/>
<Sync time="208.458"/>
il se passe des fois de ces trucs
</Turn>
<Turn speaker="spk1" startTime="209.627" endTime="237.148">
<Sync time="209.627"/>
ah oui oh mais c'est pour ça nous au moins nous on les a avertit sans les choquer

<Sync time="213.63"/>
au stade où ils en étaient chacun
<Sync time="215.519"/>
et <ENT type="time.date.rel">l'autre jour</ENT> avec <DE type="identity.name"><ENT type="pers.hum">Hervé</ENT></DE>
<Sync time="217.193"/>
et <ENT type="time.date.abs">ce soir</ENT> je fais ma petite ronde <ENT type="time.date.abs">le soir</ENT> je crois que
<Event desc="pi" type="pronounce" extent="instantaneous"/>
justement appuyer là dessus hein
<Sync time="222.368"/>
au point de vue par exemple homosexualité
<Sync time="224.209"/>
<Sync time="224.614"/>
il est difficile je trouve de
<Sync time="226.284"/>
carément quand un enfant <ENT type="time.date.rel">douze treize ans</ENT>
<Sync time="228.906"/>
on peut déjà l'ammorcer là dessus mais on peut pas moi je suis pas comme
<Event desc="pi" type="pronounce" extent="instantaneous"/>
<Sync time="233.156"/>
dire euh il faut tout leurs apprendre n'importe comment et n'importe quand
<Sync time="236.452"/>
ça non
</Turn>
<Turn speaker="spk1 spk3" startTime="237.148" endTime="240.511">
<Sync time="237.148"/>
<Who nb="1"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
<Who nb="2"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
</Turn>
<Turn speaker="spk3" startTime="240.511" endTime="259.594">
<Sync time="240.511"/>
j'ai appris des choses qui s'étaient passées <ENT type="time.date.rel">l'année dernière</ENT> au camps de pionniers de <DE type="identity.name"><ENT type="pers.hum">Benoît</ENT></DE>
<Sync time="243.32"/>
je l'ai ai appris par d'autres et <DE type="identity.name"><ENT type="pers.hum">Benoît</ENT></DE> il était victime d'un
<Sync time="245.943"/>
truc qui s'est passé hein
<Sync time="247.422"/>
des histoires de bizutage il m'en avait jamais parlé il y en a un
<Event desc="pi" type="pronounce" extent="instantaneous"/>
ce qu'on lui a fait
<Sync time="252.029"/>
<Sync time="252.459"/>
il dit moi je vous le dirais pas
<Sync time="253.322"/>

<Sync time="253.656"/>
alors sur ce <ENT type="time.date.rel">cette année</ENT>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
on a discuté de ça avec <DE type="identity.name"><ENT type="pers.hum">Benoît</ENT></DE> hein
<Sync time="257.996"/>
et ben vous savez hein
</Turn>
<Turn speaker="spk1" startTime="259.594" endTime="260.906">
<Sync time="259.594"/>
enfin <ENT type="pers.hum.tit">madame</ENT> est-ce que
<Event desc="pi" type="pronounce" extent="instantaneous"/>
</Turn>
<Turn speaker="spk3" startTime="260.906" endTime="265.823">
<Sync time="260.906"/>
j'en ai parlé à <DE type="identity.name"><ENT type="pers.hum">Jean-Baptiste</ENT></DE> moi exprès
<Sync time="262.27"/>
<Sync time="262.723"/>
à vrai dire je sais que ça se passe dans les rangers aussi
<Sync time="264.559"/>
</Turn>
<Turn speaker="spk1" startTime="265.823" endTime="267.76">
<Sync time="265.823"/>
oh oui moi je fais
<Sync time="267.04"/>
</Turn>
<Turn speaker="spk1 spk3" startTime="267.76" endTime="270.473">
<Sync time="267.76"/>
<Who nb="1"/>
et vous voyez les non
<Who nb="2"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
oui ça c'est crû
</Turn>
<Turn speaker="spk2" startTime="270.473" endTime="286.213">
<Sync time="270.473"/>
un euh un des miens euh
<Sync time="271.832"/>
<Sync time="272.381"/>
euh je crois que c'était à une colonie de vacance je sais même pas si c'était pas un camps
<Event desc="pi" type="pronounce" extent="instantaneous"/>
<Sync time="275.343"/>
<Sync time="276.278"/>
euh les grands étaient venus pou- avec des petits
<Sync time="278.668"/>
et ils avaient ils les avaient tous déshabillé et ils mesuraient leurs leurs sexes
<Sync time="282.226"/>
<Sync time="283.49"/>
pour voir lequel avait le plus gros ou le plus long
<Sync time="285.259"/>

</Turn>
<Turn speaker="spk1" startTime="286.213" endTime="300.722">
<Sync time="286.213"/>
oh oui oh mais non mais ça
<Sync time="287.644"/>
alors on peut pas laisser partir des enfants
<Sync time="289.79"/>
mais alors par contre au camps rangers <ENT type="time.date.rel">l'année dernière</ENT> y avait rien eu
<Sync time="293.038"/>
ça j'avais amorcé ah oui ah j'ai dis mon vieux hein
<Sync time="296.043"/>
dans un cas pareil ça autant on vous a
<Sync time="298.48"/>
prêché la tolérance là tu vas être
</Turn>
<Turn speaker="spk2 spk1" startTime="300.722" endTime="305.234">
<Sync time="300.722"/>
<Who nb="1"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/>
mais moi c'est ce que j'ai dis à <DE type="identity.name"><ENT type="pers.hum">Jean-Baptiste</ENT></DE> si y a n'importe quoi qui se passe il faut être assez fort pour qui disent aux autres vous êtes
<Who nb="2"/>
<Event desc="pi" type="pronounce" extent="instantaneous"/> ah
</Turn>
<Turn speaker="spk2" startTime="305.234" endTime="307.147">
<Sync time="305.234"/>
des salopards et des saligaults on fait pas des choses comme ça
</Turn>

Annexe 12 : Conventions de l'annotation multidimensionnelle des reformulations dans les transcriptions de l'oral

L'annotation se fait sous format XML.
Eléments du document XML

- **Marqueur :** <MRP>
- **Catégories syntaxiques des segments paraphrasés :** <N>, <A> etc.

catégorie seule : *N, A, V, Prep, Pr* (pronom), *PRES* (présentateur : *il y a, c'est, voici*, etc), *P* (phrase/proposition), *Adv*

catégorie avec des modifieurs : *NP, VP, AP, PP*
*Différence entre N/NP : *N* pour les noms sans déterminants et *NP* pour tout le reste, pareil pour les autres catégories syntaxiques

* La catégorie syntaxique annotée est suivie d'un numéro :

1 - pour indiquer qu'il s'agit d'un segment source qui va être paraphrasé, il se trouve avant le marqueur

2 - pour indiquer qu'il s'agit d'un segment paraphrasé qui se trouve après le marqueur

Attributs du document XML

Les attributs se mettent sur le deuxième segment, c'est-à-dire que c'est la catégorie syntaxique du deuxième segment qui possède les attributs.

- **rel_lex="hypero / syno / hypo / ant / instance / meron (mot1/mot2) "**

L'attribut rel_lex montre les relations que peuvent entretenir les unités lexicales d'un segment 1 (S1) avec segment 2 (S2). Les valeurs possibles sont :

- **hypero** : hyperonymie lorsqu'un mot ou une séquence de mots d'un S1 est un hyperonyme d'un mot ou d'une séquence de mots d'un S2. IL s'agit des relations hiérarchiques du type : sorte_de
- **hypo** : hyponymie, l'inverse du cas précédent, lorsqu'un mot ou une séquence de mots d'un S1 est un hyponyme d'un mot ou d'une séquence de mots d'un S2.
- **syno** : synonymie, un mot ou une séquence de mots d'un S1/S2 est un synonyme d'un mot ou d'une séquence de mots d'un S2/S1

*Parfois, le trait de synonymie entre les deux unités lexicales n'est pas évident mais on le met quand même car c'est comme cela qu'il est perçu par le locuteur :

*<AP1>basée euh sur le capitalisme</AP1> enfin la société française <MDR>disons</MDR> euh euh <AP2 rel_lex="**syno(capitalisme/ valeurs erronnées**)" modif_lex="remp(capitalisme/valeurs erronnées)" rel_pragm="prec">basée sur les valeurs euh euh erronées</AP2>*

- **anto** : antonymie lorsqu'un mot ou une séquence de mots d'un S1/S2 est un antonyme d'un mot ou d'une séquence de mots d'un S2/S1
- **mero** : meronymie, lorsqu'un mot ou une séquence de mots d'un S1/S2 est un meronyme d'un mot ou d'une séquence de mots d'un S2/S1. Généralement il s'agit d'une relation partie_de mais nous y avons ajouté une relation plus large fondée sur les associations que les mots peuvent avoir entre eux
- **instance** : lorsqu'un mot ou une séquence de mots d'un S1/S2 est une instance d'un mot ou d'une séquence de mots d'un S2/S1. Il s'agit souvent des entités nommées (noms propres ou groupes nominaux) qui permettent de référencer un élément

- **modif_lex="remplacement (mot1/mot2) / suppression (mot) / ajout (mot) "**

L'attribut modif_lex montre les modifications qui se font entre les unités lexicales pour passer d'un segment 1 (S1) au segment 2 (S2). Les valeurs possibles sont :

  - **remplacement** : lorsqu'un mot ou une séquence de mots d'un S1 a été remplacé par un autre
  - **ajout** : lorsqu'un mot ou une séquence de mots a été ajouté au S2
  - **suppression** : lorsqu'un mot ou une séquence de mots d'un S1 a été supprimé

- **rel_pragm="def/explic/exempl/prec/denom/res/cor_ling/paraph/justif/co_ref"**

L'attribut rel_pragm montre les relations que peuvent entretenir les unités lexicales d'un segment 1 (S1) avec le segment 2 (S2) au niveau pragmatique. Il s'agit d'une fonction pragmatique qui répond à la question pour quelle raison le locuteur fait le recours à une reformulation. Les valeurs possibles sont :

  - **def** : définition, lorsqu'un terme (simple ou composé) dans un S1 est défini dans un S2. Il s'agit souvent des termes spécialisés du domaine.
  - **explic** : explication, lorsque le locuteur essaie d'expliquer qqch à son interlocuteur. Pour vérifier cette fonction, on peut remplacer le marqueur par *parce que*
  - **justific** : justification. Cette fonction ressemble à la précédente mais dans ce cas, le locuteur présuppose que ce qu'il a dit dans S1 peut être mal perçu par son interlocuteur, il doit donc le justifier. On peut remplacer le marqueur, dans certains cas, par *c'est pourquoi*
  - **prec** : précision. Une fonction assez large dont les cas sont nombreux. Il s'agit d'une envie du locuteur d'ajouter une information dans le but de l'éclaircir.
  - **res** : résultat. Le locuteur résume le contenu du S1 ou donne en abrégé une conséquence de ce qui a été dit dans le S1. Le marqueur peut être remplacé dans ce cas pour *en somme, en bref.*
  - **exempl** : exemplification, lorsque le locuteur donne un exemple dans un S2 d'une entité mentionné dans le S1. Il peut s'agit d'une entité nommé éventuellement.
  - **dénom** : dénomination. C'est le cas qui ressemble au précédent. La différence est que contrairement à l'exemple où l'existence des autres entités du même type est présupposée, ici la fonction consiste dans l'attribution d'un nom à une entité unique.
  - **para** : paraphrase lorsqu'on ne constate aucune différente sémantique entre les deux segments.

*Les relations pragmatiques (rel_pragm) doivent être toujours annotées.

- **modif_morph="flex/deriv"**

L'attribut modif_morph montre les modifications ou les liens que les unités lexicales d'un segment 1 (S1) et d'un segment 2 (S2) entretiennent au niveau morphologique. Les valeurs possibles sont :

  - **flex** : flexion, lorsqu'un mot d'un S1 reapparaît dans un S2 sous une autre forme fléchie.
  - **deriv** : dérivation, lorsque les S1 et S2 contiennent des mots de la même famille morphologique

- **modif_synt="passif/actif"**

L'attribut modif_synt montre les modifications qui se font entre les S1 et S2 au niveau syntaxique. Il s'agit (pour l'instant) que d'une type de modification passif/actif ou actif/passif. Une seule valeur possible est :

  - passif/actif

Annexe 13 : Trois recettes de la préparation d'omelettes

***Recettes faciles* (2008) de Françoise Bernard**

Temps de préparation : 5 min

Temps de cuisson : 6 min

Ingrédients :

9 œufs

50 g de beurre

¼ de verre d'eau

sel, poivre

Nombre de personnes : 6

Préparation :

Cassez les œufs dans un petit saladier. Salez, poivrez. Versez l'eau. Battez vivement le mélange sans insister trop longtemps. Mettez le beurre dans une grande poêle. Lorsqu'il commence à roussir, versez-y les œufs battus. Détachez les bords de l'omelette en passant une fourchette entre les œufs et la poêle, tout autour. De temps en temps ramenez le bord de l'omelette vers le centre avec la fourchette. Laissez cuire cinq à six minutes. Soulevez le bord : il doit être doré tandis que le dessus de l'omelette est encore baveux. Inclinez la poêle au-dessus d'un plat et à l'aide d'une fourchette roulez l'omelette et faites-la glisser sur le plat. Servez très chaud.

Recette 2 (R2)

www.marmiton.org

Préparation : 5 min

Cuisson : 10 min

Ingrédients (pour 4 personnes) :

7 œufs

50 g de beurre

sel, poivre

Préparation :

Battez les œufs à la fourchette, salez et poivrez. Faites chauffer le beurre, versez-en un peu dans les œufs et mélangez. Versez les œufs dans la poêle à feu vif, baissez le feu et laissez cuire doucement en ramenant les bords de l'omelette au centre au fur et à mesure qu'ils prennent. Secouez un peu la poêle pour éviter que l'omelette n'attache, vérifiez la texture baveuse ou bien prise. Pliez l'omelette en deux et servez.

Recette 3 (R3)

http://recettessimples.fr/news/omelette-nature

Ingrédients :

4 œufs

Sel

Poivre

Beurre

Préparation :

Dans un bol, casser les œufs. Ajouter du sel, du poivre, battre et mélanger avec une fourchette.

Cuisson :

Dans une poêle adaptée, faire fondre un morceau de beurre. Lorsque le beurre est fondu, ajouter les œufs battus. Remuer au début avec une cuiller en bois. Lorsque l'omelette est presque cuite, mais encore « baveuse », replier une moitié sur l'autre dans la poêle. Servir

Annexe 14 : Classement de l'ordre des actions élémentaires dans les recettes d'omelettes par code et par fréquence

Code	Fréquence	Code	Fréquence
01	67	01	67
02	13	60	48
03	1	12	43
04	13	56	41
05	0	25	31
06	0	45	31
10	2	24	30
12	43	13	26
13	26	32	25
14	5	23	21
15	1	50	21
16	0	02	13
20	9	04	13
21	1	35	13
23	21	34	11
24	30	42	10
25	31	20	9
26	3	30	9
30	9	41	9
31	0	46	7
32	25	53	7
34	11	14	5
35	13	40	5
36	2	54	5
40	5	43	4
41	9	26	3
42	10	10	2
43	4	36	2
45	31	63	2
46	7	03	1
50	21	15	1
51	0	16	1

52	0		21	1
53	7		62	1
54	5		05	0
56	41		06	0
60	48		31	0
61	0		51	0
62	1		52	0
63	2		61	0
64	0		64	0
65	0		65	0

Annexe 15 : Extrait du corpus des titres de cartes géographiques annoté en lieux.

CONCA ERBAJU.
{NOTRE{ MOULIN,.LieuHistorique},.LieuGenerique+LieuSubj} Jean-Yves et Sandra.
CAGNA _ A TIA MARTINETTI Dominique.
CIAMANNACCE {et ses environs,.eTenvirons}.
CIAMANNACCE {et ses environs,.eTenvirons}.
ALTA ROCCA MARTINETTI Dominique.
BAVELLA QUERCITELLA DIGUE _ PINSON.
CENTURI Pilou & Lala.
BTP {LE CONQUET,le conquet.ChefLieuE2} Gendarme RICHARD J-L.
PORTSALL 2007 vacances de la famille toutain.
{PLABENNEC,plabennec.ChefLieuE2}.
{PLABENNEC,plabennec.ChefLieuE2}.
PLOUGASTEL PRESQU {ILE,.LieuNaturel}.
{LOCTUDY,loctudy.ChefLieuE2}.
{CAST,cast.ChefLieuE2} La {{ville,.LieuTerritoire} de votre Bonheur,.LieuGenerique+LieuSubj}.
LE {PAYS DE JAFFRES,.LieuTerritoire}.
VEUX TU M EPOUSER.
{TREGASTEL,trégastel.ChefLieuE2}.
{TERRITOIRE,.LieuTerritoire} {INZINZAC,inzinzac.ChefLieuE2} _ LOCHRIST 56.
COATCREN Les {bois de {Lochrist,fontaine de lochrist.HydronymeE2},.LieuNaturel}
{Ploerdut,ploërdut.ChefLieuE2}.
TRESTEL 2007.
{STE HELENE,sainte-hélène.ChefLieuE2} {Locoal,anse de locoal.HydronymeE2} {Nostang,nostang.ChefLieuE2}
{Landevant,landévant.ChefLieuE2} {Kervignac,kervignac.ChefLieuE3}.
RIA {ETEL,étel.ChefLieuE2} {LOCOAL,anse de locoal.HydronymeE2} {MENDON,mendon.ChefLieuE2}.
RANDONNEES VTC Françoise _ Pierrefé.
{PLOUGRESCANT,plougrescant.ChefLieuE2}.
{AUTOUR DU,.AutourDe} LATZ.
{ST BARTHELEMY,saint-barthélemy.ChefLieuE2} {Canton de {BAUD,tunnel de
baud.ToponymeCommunicationE2},.LieuTerritoire} {MORBIHAN,Morbihan.DepartementE3}.
{MALGUENAC,malguénac.ChefLieuE2} SPECIALE RANDONNEURS.
LE {PARADIS,.LieuSubj} EN FRANCE.
{ILE-DE-BREHAT,île-de-bréhat.ChefLieuE2} Michel AUDOUX.
G DU {MORBIHAN,golfe du morbihan.HydronymeE2} 50 ans Patrick Myriam et Regis.
{ARRADON,arradon.ChefLieuE2} Vacances été 2007 Manuela et Pierre.
{SOCIETE,.LieueNtreprise} CHASSE DE KERVET.
PENVINS {Centre du Monde,.LieuSubj} pour Jean-Yves.
{TAMALOU-LAND,.LieuSubjLand} jamais de pluie foi de JEANNOT.
{GUERANDE,guérande.ChefLieuE2} La {maison de Chon,.LieuBatimentHum}.
{LA TURBALLE,la turballe.ChefLieuE2} Les {marais salants,.LieuNaturel}.
SERENTAISE Foulées-VVT-{Marche,marche.RegionNaturelleE3} Trail et Randos.
{ILE,.LieuNaturel} DIEU La carte à Bruno.
LA CARTE PERSO DE LAURENT.
{CHEZ DESSUS,.LieuSubj} petites randonnées {autour de,.AutourDe} Kerhaut.
LES PEUPLIERS Jacqueline et Michel.
{SAINT-NAZAIRE,saint-nazaire.ChefLieuE2} ENF {Groupe de Gavy,.LieueNtreprise}.

YOLANDE 50 ans après toujours bretonne.
D AIGNAN.
{BARRAGE,.LieuNaturel} LURBERRIA.
NOTRE {MAISON DE CAMPAGNE,.LieuBatimentHum}.
TEAM U {NANTES,nantes.ChefLieuE3} ATLANTIQUE.
MILAFRANKA {Mon{ village,.LieuTerritoire},.LieuGenerique+LieuSubj}{ Maison,.LieuBatimentHum} MATHILDE.
{BARTHES,barthes du bourg.LieuDitNonHabiteE2} {ADOUR,adour.FleuveRiviereE3}.
URKETA {Mon{ village,.LieuTerritoire},.LieuGenerique+LieuSubj}{ Maison,.LieuBatimentHum} JONATHAN.
LES {{CHEMINS,.LieuChemin} DE CHRISTINE,.LieuGenerique+LieuSubj}.
VTT-RANDOS {Siouville-Hague,siouville-hague.ChefLieuE2}.
{VENELLES,.LieuChemin} ET {RUELLES,.LieuChemin}.
CHICHOT La carte pour {ne,né.FleuveRiviereE3} plus se perdre.
{SUCE SUR ERDRE,sucé-sur-erdre.ChefLieuE2} {et ses environs,.eTenvirons}.
{LA TRANCHE,la tranche-sur-mer.ChefLieuE2} De MT et JP De FLEVILLE.
PCS plan communal de sauvegarde.
A LA {MAISON,.LieuBatimentHum} on {ne,né.FleuveRiviereE3} sort pas sans sa carte.
PCS plan {communal,pont du communal.ToponymeCommunicationE2} de sauvegarde.
{ILE AUX OISEAUX,.LieuNaturel}.
SE RETROUVER SCI Alpha 92 Immo.
{OUDON,oudon.ChefLieuE2} {et ses environs,.eTenvirons}.
{PARENTIS,parentis-en-born.ChefLieuE2} Carte pour promenade dans les {Landes,Landes.DepartementE3}.
{POUILLON,pouillon.ChefLieuE2}.
{LA ROCHELLE,la rochelle.ChefLieuE2} {Tasdon,marais de tasdon.HydronymeE2}.
PEDANGOU {PROPRIETE,.LieuTerritoire} FAMILLE PERNOT.
{YCHOUX,ychoux.ChefLieuE2} {Le Moulin,le moulin.LieuDitNonHabiteE3}.
{CHEZ VINCENT ET PASCALINE,.LieuSubj}.
{RION DES LANDES,rion-des-landes.ChefLieuE2}.
LA BOUVRAIE {commune de {Vritz,vritz.ChefLieuE2},.LieuTerritoire} {Loire Atlantique,Loire-Atlantique.DepartementE3}.
{ST SORNIN,saint-sornin.ChefLieuE2} {Près de,.AutourDe}{ chez moi,.LieuSubj}.
PROMENADES Ouest de {St-Médard en Jalles,saint-médard-en-jalles.ChefLieuE2}.
{BORDS,bords.ChefLieuE2} GAY Michel.
ARLETTE à la {montagne,.LieuNaturel}.
{MADELEINE,madeleine.MontagneE3} à la {montagne,.LieuNaturel}.
MADO-MAURICE à la {montagne,.LieuNaturel}.
MICHOU à la {montagne,.LieuNaturel}.
MIMI à la {montagne,.LieuNaturel}.
{CESTAS,cestas.ChefLieuE2} pique-nique.
CENTRALE SHIS {HAUT,bois de haut.ToponymeDiversE2} {LIEU PYRENEEN,.LieuTerritoire}.
ESTIALESQ {Pays des TRESMONTAN,.LieuTerritoire}.
{LA TARDIERE,la tardière.ChefLieuE2} La Grande Cantière.
DE IRLEAU A {DAMVIX,damvix.ChefLieuE2}.
LE VALENTIN Randonnées entre {Ayous,pont d'ayous.ToponymeCommunicationE2} et OSSAU.
{BROCAS,brocas.ChefLieuE2} 2007.
CHOLET {Forêt de {Nuaillé,nuaillé.ChefLieuE2},.LieuNaturel}.
LA CARTE DE MAMIE DO.
CARTE PERSO {NIEUL,nieul-sur-l'autise.ChefLieuE2} SUR LAUTISE.
La carte à Papy.
MANCINI {ZONE,.LieuTerritoire} MONTOISE.

Annexe 16 : Analyse statistique effectuée par Cordial du corpus des titres de cartes géographiques extraits durant l'année 2007

TOTAUX

Nombre total de caractères, y compris les espaces et ponctuations : 111 379

Nombre total de caractères, y compris les ponctuations mais sans les espaces et tabulations : 99 366

Nombre total de mots : 15 813

Nombre total de phrases : 3 955

Nombre total de phrases verbales : 153

Nombre total de phrases de dialogue : 0

Nombre total de paragraphes : 3 388

Nombre total de paragraphes de dialogue : 0

Nombre total de ponctuations : 4 076

Nombre total de noms : 10 782

Nombre total d'adjectifs : 486

Nombre total d'adverbes : 262

Nombre total de verbes : 171

Nombre total de pronoms : 145

MOYENNES

Nombre moyen de lettres par mot : 5,41

Nombre moyen de mots par période (entre deux ponctuations, fortes ou faibles) : 3,88

Nombre moyen de mots par proposition : 3,98

Nombre moyen de mots par phrase : 4,00

Nombre moyen de mots par phrase de dialogue : 0,00

Nombre moyen de mots par phrase de non-dialogue : 4,00

Nombre moyen de mots par paragraphe : 4,67

Nombre moyen de propositions par phrase : 1,17

Nombre moyen de phrases par paragraphe : 0,00

Nombre moyen de phrases par paragraphe de dialogue : 1,17

Nombre moyen de phrases par paragraphe de non-dialogue : 1,00

MOYENNES GRAMMATICALES

Nombre moyen de substantifs par proposition : 2,72

Nombre moyen d'adjectifs par proposition : 0,12

Nombre moyen d’adverbes par proposition : 0,07

Nombre moyen de pronoms par proposition : 0,04

PHRASES

% des phrases de dialogue par rapport à l’ensemble des phrases : 0,00

% des phrases comportant au moins une proportion subordonnée par rapport à l’ensemble des phrases : 0,05

% de phrases interrogatives par rapport à l'ensemble des phrases : 0,10

% de phrases exclamatives par rapport à l'ensemble des phrases : 0,03

% de paragraphes de dialogue par rapport à l’ensemble des paragraphes : 0,00

PONCTUATIONS

% de points par rapport à l'ensemble des ponctuations : 82,02

% de points de suspension par rapport à l'ensemble des ponctuations : 0,00

% de points d'exclamation par rapport à l'ensemble des ponctuations : 0,02

% de points d'interrogation par rapport à l'ensemble des ponctuations : 0,10

% de points virgules par rapport à l'ensemble des ponctuations : 14,84

% de deux-points par rapport à l'ensemble des ponctuations : 0,00

% de virgules par rapport à l'ensemble des ponctuations : 0,15

% de parenthèses par rapport à l'ensemble des ponctuations : 0,00

% de tirets par rapport à l'ensemble des ponctuations : 2,80

% de crochets et accolades par rapport à l'ensemble des ponctuations : 0,00

MORPHOLOGIES

% de mots ambigus grammaticalement : 48,77

% de mots non ambigus grammaticalement : 100,00

% de déterminants par rapport à l'ensemble des mots : 12,52

% d'articles définis par rapport à l'ensemble des mots : 8,95

% d'articles indéfinis par rapport à l'ensemble des mots : 0,25

% d'adjectifs démonstratifs par rapport à l'ensemble des mots : 0,01

% d'adjectifs indéfinis par rapport à l'ensemble des mots : 1,40

% d'adjectifs possessifs par rapport à l'ensemble des mots : 0,03

% de prépositions par rapport à l'ensemble des mots : 8,66

% de conjonction de coordination par rapport à l'ensemble des mots : 3,08

% de conjonction de subordination par rapport à l'ensemble des mots : 0,04

% de pronoms relatifs par rapport à l'ensemble des mots : 0,06

% de pronoms possessifs par rapport à l'ensemble des mots : 0,00

% de pronoms personnels à la 1e personne du singulier par rapport à l'ensemble des pronoms possessifs : 23,42

% de pronoms personnels à la 2e personne du singulier par rapport à l'ensemble des pronoms possessifs : 3,15

% de pronoms personnels à la 3e personne du singulier par rapport à l'ensemble des pronoms possessifs : 36,49

% de pronoms personnels à la 1e personne du pluriel par rapport à l'ensemble des pronoms possessifs : 28,38

% de pronoms personnels à la 2e personne du pluriel par rapport à l'ensemble des pronoms possessifs : 8,11

% de pronoms personnels à la 3e personne du pluriel par rapport à l'ensemble des pronoms possessifs : 0,45

% de mots-outils par rapport à l'ensemble des mots : 26,00

% de mots signifiants (c.à-d. non mots-outils) par rapport à l'ensemble des mots : 74,00

% de substantifs par rapport au total des mots signifiants (substantifs, adjectifs, verbes, adverbes) : 92,15

% d'adjectifs par rapport au total des mots signifiants (substantifs, adjectifs, verbes, adverbes) : 4,15

% de verbes par rapport au total des mots signifiants (substantifs, adjectifs, verbes, adverbes) : 1,46

% d'adverbes par rapport au total des mots signifiants (substantifs, adjectifs, verbes, adverbes) : 2,24

USAGE

% de mots appartenant au vocabulaire de base (Gougenheim) par rapport à l'ensemble des mots : 40,59

% de mots appartenant à la liste Dubois-Buyse par rapport à l'ensemble des mots : 39,25

% de mots très courants par rapport à l'ensemble des mots : 23,20

% de mots courants par rapport à l'ensemble des mots : 67,41

% de mots rares par rapport à l'ensemble des mots : 9,39

% de mots très rares par rapport à l'ensemble des mots : 6,70

% de mots inconnus par rapport à l'ensemble des mots : 22,51

% de noms communs inconnus par rapport à l'ensemble des noms communs : 17,18

% de noms propres inconnus par rapport à l'ensemble des noms propres : 5,33

NOMS

% de noms communs par rapport aux noms (communs ou propres) : 71,91

% de noms propres par rapport aux noms (communs ou propres) : 28,09

% de noms composés par rapport aux noms (communs ou propres) : 3,63

% de noms abstraits par rapport à l'ensemble des noms : 46,15

% de noms concrets par rapport à l'ensemble des noms : 53,85

NOMS COMMUNS

% de noms communs ayant des homonymes parmi l'ensemble des substantifs : 42,55

% de noms communs ayant des paronymes parmi l'ensemble des substantifs : 63,65

% de noms communs de lieu parmi l'ensemble des substantifs : 35,40

% de noms communs d'adressage parmi l'ensemble des substantifs : 6,42

% de noms communs de temps parmi l'ensemble des substantifs : 6,74

% de noms communs de nationalité parmi l'ensemble des substantifs : 0,06

% de noms communs de profession parmi l'ensemble des substantifs : 1,61

% de noms communs de type uniquement humain parmi l'ensemble des substantifs : 5,41

% de noms communs de type uniquement humanoïde parmi l'ensemble des substantifs : 0,03

% de noms communs de type uniquement animal parmi l'ensemble des substantifs : 1,93

% de noms communs de type uniquement animé parmi l'ensemble des substantifs : 0,13

% de noms communs de type uniquement concret parmi l'ensemble des substantifs : 51,09

% de noms communs de type uniquement abstrait parmi l'ensemble des substantifs : 20,78

% de noms communs de type pouvant être humain parmi l'ensemble des substantifs : 10,50

% de noms communs de type pouvant être humanoïde parmi l'ensemble des substantifs : 0,44

% de noms communs de type pouvant être animal parmi l'ensemble des substantifs : 2,37

% de noms communs de type pouvant être animé parmi l'ensemble des substantifs : 1,08

% de noms communs de type pouvant être concret parmi l'ensemble des substantifs : 64,82

% de noms communs de type pouvant être abstrait parmi l'ensemble des substantifs : 30,62

% de noms communs de type collectif parmi l'ensemble des substantifs : 4,21

% de noms communs monosémiques parmi l'ensemble des substantifs : 36,54

% de mots signifiants polysémiques parmi l'ensemble des mots polysémiques : 20,20

% de noms communs épithètes parmi l'ensemble des substantifs : 2,05

% de noms communs appartenant à des sous-groupes nominaux parmi l'ensemble des substantifs : 1,42

NOMS PROPRES

% d'abréviations parmi l'ensemble des noms propres : 10,56

% de noms propres de type humain parmi l'ensemble des noms propres : 58,44

% de noms propres de type prénom parmi l'ensemble des noms propres : 24,50

% de noms propres de type géographique parmi l'ensemble des noms propres : 48,60

% de noms propres de type ni humain ni géographique parmi l'ensemble des noms propres : 100,00

DETERMINANTS

% d'article définis par rapport à l'ensemble des déterminants : 97,25

% d'article indéfinis par rapport à l'ensemble des déterminants : 2,75

% d'articles définis par rapport à l'ensemble des articles : 71,50

% d'adjectifs démonstratifs par rapport à l'ensemble des déterminants : 0,10

% d'adjectifs cardinaux par rapport à l'ensemble des déterminants : 12,94

% d'adjectifs interrogatifs par rapport à l'ensemble des déterminants : 0,00

% d'adjectifs indéfinis par rapport à l'ensemble des déterminants : 0,20

% d'adjectifs possessifs par rapport à l'ensemble des déterminants : 11,22

% d'adjectifs possessifs à la 1e personne du singulier par rapport à l'ensemble des adjectifs possessifs : 0,00

% d'adjectifs possessifs à la 2e personne du singulier par rapport à l'ensemble des adjectifs possessifs : 0,00

% d'adjectifs possessifs à la 3e personne du singulier par rapport à l'ensemble des adjectifs possessifs : 0,00

% d'adjectifs possessifs à la 1e personne du pluriel par rapport à l'ensemble des adjectifs possessifs : 0,00

% d'adjectifs possessifs à la 2e personne du pluriel par rapport à l'ensemble des adjectifs possessifs : 0,00

% d'adjectifs possessifs à la 3e personne du pluriel par rapport à l'ensemble des adjectifs possessifs : 0,00

PRONOMS

% de pronoms démonstratifs par rapport à l'ensemble des pronoms : 6,21

% de pronoms indéfinis par rapport à l'ensemble des pronoms : 3,45

% de pronoms relatifs par rapport à l'ensemble des pronoms : 6,21

% de pronoms personnels par rapport à l'ensemble des pronoms : 84,14

% de pronoms possessifs par rapport à l'ensemble des pronoms : 0,00

% de pronoms personnels parmi l'ensemble des sujets : 14,38

% de pronoms sujets parmi l'ensemble des pronoms : 18,95

% de pronoms personnels à la 1e personne du singulier parmi l'ensemble des pronoms personnels sujets : 13,93

% de pronoms personnels à la 2e personne du singulier parmi l'ensemble des pronoms personnels sujets : 13,93

% de pronoms personnels à la 1e personne du pluriel parmi l'ensemble des pronoms personnels sujets : 30,33

% de pronoms personnels à la 2e personne du pluriel parmi l'ensemble des pronoms personnels sujets : 5,74

% de pronoms personnels à la 1e personne du singulier parmi l'ensemble des pronoms personnels sujets dans les dialogues : 0,00

% de pronoms personnels à la 2e personne du singulier parmi l'ensemble des pronoms personnels sujets dans les dialogues : 0,00

% de pronoms personnels à la 1e personne du pluriel parmi l'ensemble des pronoms personnels sujets dans les dialogues : 31,82

% de pronoms personnels à la 2e personne du pluriel parmi l'ensemble des pronoms personnels sujets dans les dialogues : 9,09

ADJECTIFS

% d'adjectifs ayant des homonymes parmi l'ensemble des adjectifs : 18,90

% d'adjectifs ayant des paronymes parmi l'ensemble des adjectifs : 55,34

% d'adjectifs toujours utilisés avec un substantif humain, humanoïde ou animal parmi l'ensemble des adjectifs : 2,47

% d'adjectifs toujours utilisés avec un substantif animé parmi l'ensemble des adjectifs : 0,00

% d'adjectifs toujours utilisés avec un substantif concret parmi l'ensemble des adjectifs : 5,75

% d'adjectifs toujours utilisés avec un substantif abstrait parmi l'ensemble des adjectifs : 15,07

% d'adjectifs utilisables avec un substantif humain, humanoïde ou animal parmi l'ensemble des adjectifs : 49,59

% d'adjectifs utilisables avec un substantif animé parmi l'ensemble des adjectifs : 36,71

% d'adjectifs utilisables avec un substantif concret parmi l'ensemble des adjectifs : 59,73

% d'adjectifs utilisables avec un substantif abstrait parmi l'ensemble des adjectifs : 66,30

% d'adjectifs de lieu parmi l'ensemble des adjectifs : 1,37

% d'adjectifs de temps parmi l'ensemble des adjectifs : 4,93

% d'adjectifs de couleur parmi l'ensemble des adjectifs : 4,11

% d'adjectifs épithètes par rapport à l'ensemble des adjectifs : 100,00

% d'adjectifs antéposés par rapport à l'ensemble des adjectifs préposés et postposés : 25,00

ADVERBES

% d'adverbes ayant des homonymes parmi l'ensemble des adverbes : 13,36

% d'adverbes ayant des paronymes parmi l'ensemble des adverbes : 3,82

% d'adverbes de temps parmi l'ensemble des adverbes : 6,11

% d'adverbes de lieu parmi l'ensemble des adverbes : 58,02

% d'adverbes de quantité parmi l'ensemble des adverbes : 9,92

% d'adverbes de manière parmi l'ensemble des adverbes : 22,14

% d'adverbes d'affirmation parmi l'ensemble des adverbes : 2,29

% d'adverbes de négation parmi l'ensemble des adverbes : 13,36

% d'adverbes de doute parmi l'ensemble des adverbes : 0,38

% d'adverbes précédant un adjectif parmi l'ensemble des adverbes : 5,73

% d'adverbes suivant un verbe parmi l'ensemble des adverbes : 4,96

VERBES

% de verbes au présent par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 76,05

% de verbes à l'imparfait par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 1,20

% de verbes au passé simple par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 2,99

% de verbes au futur par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 0,60

% de verbes au conditionnel par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 0,00

% de verbes au subjonctif présent par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 5,39

% de verbes au subjonctif imparfait par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 0,00

% de verbes au subjonctif passé par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 0,00

% de verbes au subjonctif plus-que-parfait par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 0,00

% de verbes au passé composé par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 2,40

% de verbes au plus-que-parfait par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 0,00

% de verbes au passé antérieur par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 0,00

% de verbes au futur antérieur par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 0,00

% de verbes au conditionnel passé par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 0,00

% de verbes à l'impératif par rapport à l'ensemble des verbes conjugués (temps simples et composés) : 11,38

% de verbes à la 1e personne du singulier parmi l'ensemble des verbes conjugués (temps simples et composés) : 6,59

% de verbes à la 2e personne du singulier parmi l'ensemble des verbes conjugués (temps simples et composés) : 17,96

% de verbes à la 3e personne du singulier parmi l'ensemble des verbes conjugués (temps simples et composés) : 65,27

% de verbes à la 1e personne du pluriel parmi l'ensemble des verbes conjugués (temps simples et composés) : 5,39

% de verbes à la 2e personne du pluriel parmi l'ensemble des verbes conjugués (temps simples et composés) : 3,59

% de verbes à la 3e personne du pluriel parmi l'ensemble des verbes conjugués (temps simples et composés) : 1,20

% des verbes ayant des homonymes parmi l’ensemble des verbes : 49,31

% des verbes ayant des paronymes parmi l’ensemble des verbes : 65,28

% des verbes pronominaux parmi l’ensemble des verbes : 0,00

% des verbes intransitifs parmi l'ensemble des verbes : 9,72

% des verbes transitifs directs parmi l'ensemble des verbes : 56,25

% des verbes transitifs indirects parmi l'ensemble des verbes : 30,56

% des verbes prenant l'auxiliaire « avoir » parmi l'ensemble des verbes : 13,89

% des verbes prenant l'auxiliaire « être » parmi l'ensemble des verbes : 7,64

% des verbes à COD obligatoire parmi l'ensemble des verbes : 45,83

% des verbes à COI obligatoire parmi l'ensemble des verbes : 9,72

% des verbes toujours impersonnels parmi l'ensemble des verbes : 0,00

% des verbes parfois impersonnels parmi l'ensemble des verbes : 4,86

% des verbes précédés d'une négation parmi l'ensemble des verbes : 1,20

% des verbes de lieu parmi l'ensemble des verbes : 13,19

% des verbes de temps parmi l'ensemble des verbes : 4,86

% des verbes à sujet uniquement humain ou humanoïde parmi l'ensemble des verbes : 35,42

% des verbes à sujet uniquement animal parmi l'ensemble des verbes : 0,00

% des verbes à sujet uniquement animé parmi l'ensemble des verbes : 0,00

% des verbes à sujet uniquement concret parmi l'ensemble des verbes : 2,08

% des verbes à sujet uniquement abstrait parmi l'ensemble des verbes : 2,08

% des verbes à sujet pouvant être humain ou humanoïde parmi l'ensemble des verbes : 81,25

% des verbes à sujet pouvant être animal parmi l'ensemble des verbes : 40,28

% des verbes à sujet pouvant être animé parmi l'ensemble des verbes : 34,03

% des verbes à sujet pouvant être concret parmi l'ensemble des verbes : 35,42

% des verbes à sujet pouvant être abstrait parmi l'ensemble des verbes : 39,58

% des verbes à COD uniquement humain ou humanoïde parmi l'ensemble des verbes : 9,03

% des verbes à COD uniquement animal parmi l'ensemble des verbes : 0,69

% des verbes à COD uniquement animé parmi l'ensemble des verbes : 0,69

% des verbes à COD uniquement concret parmi l'ensemble des verbes : 11,81

% des verbes à COD uniquement abstrait parmi l'ensemble des verbes : 6,94

% des verbes à COD pouvant être humain ou humanoïde parmi l'ensemble des verbes : 31,25

% des verbes à COD pouvant être animal parmi l'ensemble des verbes : 19,44

% des verbes à COD pouvant être animé parmi l'ensemble des verbes : 20,83

% des verbes à COD pouvant être concret parmi l'ensemble des verbes : 34,72

% des verbes à COD pouvant être abstrait parmi l'ensemble des verbes : 32,64

% des verbes à COI uniquement humain ou humanoïde parmi l'ensemble des verbes : 11,11

% des verbes à COI uniquement animal parmi l'ensemble des verbes : 0,00

% des verbes à COI uniquement animé parmi l'ensemble des verbes : 0,00

% des verbes à COI uniquement concret parmi l'ensemble des verbes : 0,00

% des verbes à COI uniquement abstrait parmi l'ensemble des verbes : 6,94

% des verbes à COI pouvant être humain ou humanoïde parmi l'ensemble des verbes : 22,22

% des verbes à COI pouvant être animal parmi l'ensemble des verbes : 7,64

% des verbes à COI pouvant être animé parmi l'ensemble des verbes : 7,64

% des verbes à COI pouvant être concret parmi l'ensemble des verbes : 11,11

% des verbes à COI pouvant être abstrait parmi l'ensemble des verbes : 19,44

PROPOSITIONS

% de propositions indépendantes parmi l'ensemble des propositions : 99,90

% de propositions principales parmi l'ensemble des propositions : 0,00

% de propositions relatives parmi l'ensemble des propositions : 0,03

% de propositions subordonnées parmi l'ensemble des propositions : 0,08

% de propositions coordonnées parmi l'ensemble des propositions : 0,00

% de propositions participiales parmi l'ensemble des propositions : 0,00

% de propositions incises parmi l'ensemble des propositions : 0,00

% de propositions participiales parmi l'ensemble des propositions : 0,00

% de propositions ayant un complément d'objet direct : 1,64

% de propositions ayant un complément d'objet indirect : 0,53

% de propositions ayant un attribut du sujet : 0,40

% de propositions ayant un complément circonstanciel de temps : 0,13

% de propositions ayant un complément circonstanciel de lieu : 0,58

% de propositions ayant un complément circonstanciel ni de temps ni de lieu : 0,81

STYLE

% de clichés par rapport à l'ensemble des mots : 0,00

% d'expressions usuelles par rapport au nombre de mots : 0,00

degré de lisibilité : 0,36

TYPES GRAMMATICAUX

% Adjectifs masculins singuliers par rapport aux types grammaticaux : 0,78

% Adjectifs masculins pluriels par rapport aux types grammaticaux : 0,23

% Adjectifs féminins singuliers par rapport aux types grammaticaux : 0,68

% Adjectifs féminins pluriels par rapport aux types grammaticaux : 0,24

% Adjectifs masculins invariants en nombre par rapport aux types grammaticaux : 0,18

% Adjectifs féminins invariants en nombre par rapport aux types grammaticaux : 0,00

% Adjectifs singuliers invariants en genre par rapport aux types grammaticaux : 0,37

% Adjectifs pluriels invariants en genre par rapport aux types grammaticaux : 0,10

% Adjectifs invariants en genre et en nombre par rapport aux types grammaticaux : 0,49

% Adjectifs démonstratifs par rapport aux types grammaticaux : 0,01

% Adjectifs possessifs par rapport aux types grammaticaux : 1,40

% Adjectifs numériques cardinaux par rapport aux types grammaticaux : 1,78

% Adjectifs numériques ordinaux par rapport aux types grammaticaux : 0,25

% Adverbes par rapport aux types grammaticaux : 1,66

% Articles définis masculins singuliers par rapport aux types grammaticaux : 3,73

% Articles définis féminins singuliers par rapport aux types grammaticaux : 2,44

% Articles pluriels invariants en genre par rapport aux types grammaticaux : 2,78

% Articles indéfinis masculins singuliers par rapport aux types grammaticaux : 0,17

% Articles indéfinis féminins singuliers par rapport aux types grammaticaux : 0,08

% Conjonctions de coordinations par rapport aux types grammaticaux : 3,08

% Conjonctions de subordinations par rapport aux types grammaticaux : 0,04

% Interjections par rapport aux types grammaticaux : 0,11

% Prépositions par rapport aux types grammaticaux : 8,66

% Noms masculins singuliers par rapport aux types grammaticaux : 18,04

% Noms masculins pluriels par rapport aux types grammaticaux : 2,85

% Noms féminins singuliers par rapport aux types grammaticaux : 8,37

% Noms féminins pluriels par rapport aux types grammaticaux : 2,47

% Noms masculins invariants en nombre par rapport aux types grammaticaux : 3,40

% Noms féminins invariants en nombre par rapport aux types grammaticaux : 0,04

% Noms singuliers invariants en genre par rapport aux types grammaticaux : 8,18

% Noms pluriels invariants en genre par rapport aux types grammaticaux : 0,06

% Noms invariants en genre et en nombre par rapport aux types grammaticaux : 24,78

% Pronoms personnels 1e personne du singulier par rapport aux types grammaticaux : 0,11

% Pronoms personnels 2e personne du singulier par rapport aux types grammaticaux : 0,11

% Pronoms personnels 3e personne du singulier par rapport aux types grammaticaux : 0,27

% Pronoms personnels 1e personne du pluriel par rapport aux types grammaticaux : 0,23

% Pronoms personnels 2e personne du pluriel par rapport aux types grammaticaux : 0,04

% Pronoms personnels 3e personne du pluriel par rapport aux types grammaticaux : 0,01

% Pronoms démonstratifs singuliers par rapport aux types grammaticaux : 0,05

% Pronoms démonstratifs pluriels par rapport aux types grammaticaux : 0,01

% Adjectifs interrogatifs par rapport aux types grammaticaux : 0,00

% Adjectifs indéfinis par rapport aux types grammaticaux : 0,03

% Pronoms relatifs masculins singuliers par rapport aux types grammaticaux : 0,00

% Pronoms relatifs féminins singuliers par rapport aux types grammaticaux : 0,00

% Pronoms relatifs pluriels par rapport aux types grammaticaux : 0,06

% Pronoms indéfinis masculins singuliers par rapport aux types grammaticaux : 0,01

% Pronoms indéfinis féminins singuliers par rapport aux types grammaticaux : 0,00

% Pronoms indéfinis masculins pluriels par rapport aux types grammaticaux : 0,01

% Pronoms indéfinis féminins pluriels par rapport aux types grammaticaux : 0,00

% Pronoms indéfinis singuliers invariants en genre par rapport aux types grammaticaux : 0,00

% Pronoms indéfinis pluriels invariants en genre par rapport aux types grammaticaux : 0,00

% Pronoms indéfinis invariants en genre et en nombre par rapport aux types grammaticaux : 0,01

% Pronoms possessifs par rapport aux types grammaticaux : 0,00

% verbes par rapport aux types grammaticaux : 1,59

Annexe 17 : Conventions de l'annotation du corpus des titres en lieux subjectifs

Pour toutes les annotations, le déterminant ne fait pas partie de l'annotation.

Pers : Appropriation, personnalisation de lieu :

Le lieu est désigné par un nom propre ou un nom commun générique de lieu, géographique (montagne) ou non géographique (coin, maison). L'appropriation est marquée par différents procédés qui correspondent à un attribut supplémentaire *(/attribut*). Un seul attribut est possible.

1) <Pers/poss>

- L'appropriation est marquée par un possessif (déterminant ou pronom possessifs avant ou après) qui fait partie de la balise.

Exemples *: ma campagne, Ma Montagne Saint Sorlin, ma zone à moi, notre fief, notre coin, mon Paris, chez-moi, mon chez-moi, mon sud, ma maison, chez la famille X*

possessif + (LieuGenerique/LieuGazetier/PointCardinal) + à Pronom tonique

2). **<Pers/compl>**

- complément humain introduit par les prépositions *de* ou *à*

- L'appropriation est marquée par un complément désignant une personne (prénom, nom, diminutif, métier, lien de parenté) ou un animal introduit par de ou à.

Exemples *: fief de Patrick XXX, coin de Guy, village de ma Maman, fief de la famille X, PAYS DES CHTIS, pays des abeilles, coin des ours", coin de l'ours, pays à Papa, pays de Papa*

(LieuGenerique/LieuGazetier/PointCardinal) + de/à + (det) + Personne : nom, diminutif, métier, parenté)

3). **<Pers/autour>**

- construction où une personne est employée comme un lieu

- Après la préposition locative : "autour" au sens large, le nom désigne une personne (prénom, nom, diminutif, métier, lien de parenté, animal). La préposition fait partie de la balise.

- Pour l'annotation manuelle, la préposition introductive fait partie de la balise ; en cas de compléments de nom multiples, juxtaposés ou coordonnées, l'ensemble des compléments de nom sont inclus dans la balise.

Exemples *: Autour de Sylvie et Dom, environs de Edith et Luc, Sylvie et alentours*

autour de/environs de/près de/à côté de/surrounds… + (det) Personne : nom, diminutif, parenté, …

4). **<Pers/land>**

- C'est une création lexicale formée d'un mot suffixé par "land"; Le suffixe est ou non séparé du mot précédant ; différents séparateurs sont admissibles : , -, _, '.

Exemples : *Papyland, Zozo land, TAMALOU-LAND, PLOUCLAND*

5). **<Pers/autres>**

Le lieuGazetier (désignant donc un lieu en France) est accompagné d'un nom générique géographique ou administratif dans une langue étrangère et n'est pas "land". Ce nom générique peut être séparé du lieuGazetier (et situé avant ou après) ou bien accolé au nomGazetier.

Exemples : *Rouen city, CESSON Y ALREDEDORES, BLUPSCITY, POUILLART CITY*

Invente : Lieux imaginaires, inventés = lieux dans les œuvres artistiques

<Invente>

Pour l'annotation manuelle, le déterminant quand il existe ne fait pas partie de la séquence annotée.

Exemples : *Terres du Milieu, Atlantide*

Metaph : Lieux métaphoriques = désignations des lieux en comparaison aux autres lieux existants

<Metaph>

- Lieux en France comparés ou associés à un lieu étranger
- Le lieu utilisé pour la comparaison est un nom propre désignant un lieu étranger actuel ou non**.**

Exemples : *Amérique Française,*

Sentiment : Evaluation du lieu, expression d'un sentiment sur le lieu :

1). **<Sentiment/modif>**

- lieu (LieuGenerique ou LieuGazetier ou PointCardinal) accompagné d'un lexique (nom ou adjectif) de sentiment ou d'appréciation

Exemples : *lieu magique, pays préféré, paysages magiques, tanière idéale, villa aimée, bons coins, joli petit coin, ILE des rêves, maison du bonheur, plus beau château*

LieuGenerique / LieuGazetier + adj_de_sentiment (il peut y avoir plusieurs adjectifs)
(plus) + adj_de_sentiment + LieuGenerique
LieuGenerique / LieuGazetier / PointCardinal + de + nom_de_sentiment

2) <Sentiment/figé>

Le lieu est désigné par une expression figée qui ne désigne pas nécessairement un lieu ("monts et merveilles")

NB : "maison de famille" est une expression polylexicale mais pas une expression figée

Exemples : *trou du cul, centre du monde, , bout du chemin, sweet home, chemin du retour, , havre de paix, far east, terra nostra, mont et merveille, entre mer et terre*

LieuxSubjListe : Liste des lieux subjectifs composés
centre / bout / cœur / coin de + LieuxSubjListe par nature

3) <Sentiment/nature>

le lieu (nom simple ou composé) contient déjà une évaluation ou un sentiment par défaut, ce n'est pas un lieu neutre par nature parce qu'il est désigné :

- par un nom qui renvoie à un lieu qui ne peut pas correspondre à un lieu en France : paradis, royaume, univers, empire, …
- par un nom employé au sens figuré : berceau, racine, attache, …
- par un nom LieuBatimentAnimal : tanière, pigeonnier, niche, …

Exemples : *paradis, royaume, berceau, univers, empire, galaxie, monde, planète, racine, attache etc.*

ex. {PARADIS EN FRANCE}

4) <Sentiment/typo>

présence d'une information extralinguistique :
- des signes typographiques, propres ou non au Langage Web (☺, ;), !! etc.), positionnés avant ou après le nom de lieu
- et/ou d'un commentaire sur le nom de lieu.

Pour l'annotation manuelle, signes typographiques et commentaire sont inclus dans la balise.

Exemples : *"Nice Aire St Michel",*

Fonction

<Fonction>

Lieu accompagné par une mention de fonction (introduite souvent par un groupe prépositionnel en *à*)

1). Exemples : *chemins à parcourir, circuits à fleurs et à champignons, chemins à découvrir, à explorer*

2). Exemples : *là où je cours*

LieuGenerique / LieuGazetier + à + V-inf
LieuChemin + à

Aujourd'hui, on tient compte de la prépositions à ou pour ou « de »

Les séquences qui commencent par un déictique (du type : "{*là où je cours}*") ne sont pas balisées par les patrons, mais elles doivent être annotées manuellement.

Lieu unique

<Unique>

nom de lieu précédé d'un déterminant défini

exemples : *La montagne, Le Paris*

Les lieux combinés

Si une désignation relève de plusieurs cas de subjectivités, les annotations s'ajoutent. Il n'y a pas de limitation dans le nombre d'annotations combinées. Les balises correspondantes sont réunies par le signe"+" sans espace avant ni après. L'ordre des balises n'est pas significatif (de leur importance par exemple), mais doit être le suivant :

Invente, Sentiment, Pers, Fonction, Metaph, Typo

Si la balise combinée contient <Sentiment> ou <Pers>, l'attribut est aussi à préciser.

<Sentiment/modif+Pers/poss> : *ma maison idéale, mon trou du cul*

<Invente+Pers> : *mon Caamelot*

<Pers+Metaph> : *ma Suisse, mon Amérique, ma Perse*

<Sentiment+Metaph> : *Suisse d'amour*

<Sentiment+Pers+Metaph> : *ma petite Suisse*

<Invente+Sentiment+Pers>

<Metaph+Typo> *Le Caylar En colombie... Si, si!.*

<*Pers+Typo*> la maison de … Guy XXX, *maison :),*

Doute

Si l'on est pas sûr de l'annotation, on ajoute un attribut doute à la fin :

<Pers/pos/doute> etc.