La dynamique du texte, corpus, outils, analyses

Marie-Paule Jacques

LIDILEM

UNIVERSITÉ Grenoble Alpes

La dynamique du texte, corpus, outils, analyses

Dossier pour l'habilitation à diriger des recherches

Volume 1 : Synthèse des travaux

Marie-Paule Jacques

Jury :

Catherine Brissaud, Université Grenoble Alpes, examinatrice

Cécile Fabre, Université Toulouse Jean-Jaurès, rapporteure

Francis Grossmann, Université Grenoble Alpes, garant

Thierry Poibeau, LATTICE, UMR, rapporteur

Dirk Siepmann, Université d'Osnabrück, examinateur

Frédérique Sitri, Université Paris Ouest, rapporteure

5 décembre 2017

"L'important n'est pas ce qu'on fait de nous, mais ce que nous faisons nous-même de ce qu'on a fait de nous." *Saint-Genêt*, J.-P. Sartre

Remerciements

Mes premiers remerciements vont à celui qui me soutient depuis toujours, croit en moi plus que moi-même et fait que le quotidien est une aventure somme toute intéressante et rigolote...

Je remercie sincèrement les collègues qui acceptent de donner leur temps pour me lire et – j'espère – critiquer mes travaux. Merci à vous, membres du jury : Catherine Brissaud, Cécile Fabre, Francis Grossmann, qui a en outre accepté la responsabilité d'être mon « garant », Thierry Poibeau, Dirk Siepmann et Frédérique Sitri. J'espère n'avoir pas laissé trop de coquilles et d'erreurs qui vous compliqueraient la tâche.

On le verra dans ce mémoire, une recherche stimulante est fonction des lieux, laboratoires de recherche, équipes de chercheurs et d'enseignants. J'avais déjà dans ma thèse remercié les membres de l'ERSS, j'ai mesuré en écrivant cette synthèse à quel point je suis redevable à cette équipe, je la remercie encore à nouveau.

Et j'ai trouvé au Lidilem, particulièrement dans l'axe 1 « Descriptions Linguistiques, Corpus et TAL », un environnement tout aussi chaleureux et stimulant. Merci à Agnès Tutin et à Cristelle Cavalla, pour l'accueil dans le labo et dans le bureau, merci à Francis Grossmann, à Olivier Kraif, à Iva Novakova, à Laura Hartwell, arrivée peu après moi et déjà repartie vers d'autres cieux. Un merci spécial à Virginie Zampa, ma « jumelle de thèse », avec qui nous avons lancé la rubrique Varia de la revue Lidil, ce qui nous a permis d'échanger, parfois même dans la bonne humeur, nos impressions sur les soumissions que nous recevions et sur la recherche en général. Merci aux autres membres de l'axe, permanents et doctorants, qui participent à l'ambiance conviviale et studieuse du laboratoire, en particulier à Isabelle Rousset, toujours bien informée sur tous les rouages de l'université et toujours (ou presque) de bon conseil.

Merci aux collègues des autres axes du Lidilem. Jean-Pierre Chevrot, ex- et actuel directeur, Marinette Matthey, ex-directrice et qui a elle aussi partagé l'aventure Varia de Lidil, ainsi qu'aux autres acteurs du comité de rédaction de Lidil : Françoise Boch, Julie Sorba, Catherine Frier. Des remerciements particuliers pour mes collègues de l'ESPE, Catherine Brissaud et Fanny Rinck, avec lesquelles l'avenir se montre aussi prometteur que les années écoulées.

Que celles et ceux qui ne sont pas nommément cités ne se sentent pas exclus, je suis consciente de tout ce que l'on doit à un entourage qui sait contribuer à faire d'un laboratoire de recherche un lieu non seulement scientifique mais aussi convivial et humain, merci à tous et toutes.

Sommaire

Liste synthétique des corpus utilisés dans les recherches

Les corpus utilisés sont présentés en détail dans la section 1.4, la présente liste servira d'aide-mémoire.

[**Déplacements**] : textes élaborés dans un contexte professionnel par quatre acteurs de la gestion des déplacements sur l'agglomération toulousaine : Mairie de Toulouse, DDE (Direction Départementale de l'Équipement), SEMVAT (compagnie de transports en commun), SMTC (Syndicat Mixte de Transports en Commun).

[**Vol Libre**] : une vingtaine d'articles paru dans la presse spécialisée du vol libre, les deux magazines *Parapente Mag* et *Vol Libre*, dans la rubrique « essais de voiles ».

[**Bricolage**] : pages web, fiches pratiques mises à disposition par les sites Mr Bricolage et Maison Facile, pour aider le lecteur aussi bien pour des réalisations techniques (comme poser un carrelage ou repeindre un radiateur) que pour des actes du quotidien (comme régler un litige avec un voisin).

[**Médecine**] : environ 80 recommandations médicales, qui sont des textes élaborés par des médecins chercheurs à destination des praticiens en vue de diffuser auprès de ces derniers les avancées de la recherche médicale et de leur permettre d'adopter de bonnes pratiques médicales.

[**articles scientifiques – ingénierie des connaissances**] : articles scientifiques issus de la conférence Ingénierie des Connaissances, qui a lieu tous les ans en France.

[**articles scientifiques – géopolitique**] : articles scientifiques de géopolitique aspirés sur le site web de l'IFRI[1] (Institut Français des Relations Internationales).

[**Scientext**] : assemblage de textes scientifiques et académiques, en français et an anglais, constitué et mis à disposition par le LIDILEM à travers un site web[2].

[**Termith**] : certains textes collectés par le projet Scientext plus des articles scientifiques aspirés sur les sites web de revues de Sciences Humaines et Sociales, couvrant dix disciplines différentes de SHS : linguistique, psychologie, sciences de l'éducation, traitement automatique des langues, anthropologie, géographie, histoire, sciences de l'information et de la communication, sciences politiques, sociologie.

[**EIIDA**] : corpus collecté dans le cadre du projet EIIDA, Étude Interdisciplinaire et Interlinguistique du Discours Académique, de textes d'articles scientifiques et d'enregistrements de présentations de conférences dans deux domaines, linguistique et géochimie, et dans trois langues, anglais, espagnol, français.

[**Littéracie Avancée**] : textes de divers genres académiques (mémoires, dossiers, rapports de stage, etc.) rédigés en français par des étudiants dans des universités françaises, dans le cadre ordinaire de cursus de sciences du langage ou de formation d'enseignants, de la licence à la maitrise.

[**Sup-chinois**] : mémoires de master, rédigés en français par des étudiants chinois apprenant le français, dans des disciplines de sciences humaines variées : littérature, sociologie, linguistique.

1 http://www.ifri.org/ page consultée le 1er octobre 2017.

2 http://scientext.msh-alpes.fr/ page consultée le 1er octobre 2017.

Chapitre 1 - Introduction : un parcours et des choix théoriques

Ce document constitue une synthèse de travaux qui se sont échelonnés de l'année 2000, date de ma première publication de recherche, à l'année 2017, date à laquelle je rédige.

Une synthèse de travaux, c'est un genre qui n'est guère familier à un enseignant-chercheur – en tout cas qui ne m'est guère familier. Comment harmoniser compte-rendu de recherche et références aux travaux des autres ? Comment structurer ? « Mets de l'ordre », « Rends ton parcours visible » m'a-t-on conseillé. J'ai remercié des conseils mais suis restée assez perplexe. L'enjeu est là : donner à voir une cohérence interne qui ne s'était jamais verbalisée, puisque faite de choix successifs dont la logique tenait à la fois de ce qu'ils permettaient de poursuivre un chemin dans des directions pertinentes au regard du déjà-là, et de ce qu'ils répondaient à des opportunités qui s'offraient, nées de rencontres et de sollicitations au sein d'un réseau construit lors de mon travail de thèse. L'enjeu de cette synthèse est de faire d'une diversité un éclectisme (au sens philosophique du terme) et non un éparpillement.

Quelques mots sur mon parcours scolaire et professionnel, qui n'a aucunement été rectiligne, éclaireront les centres d'intérêts développés ultérieurement et aideront à comprendre comment s'entrelacent les thématiques sur lesquelles j'ai travaillé ces quinze dernières années (en fait un peu plus de quinze, si je comptabilise depuis ma première publication de linguiste en 2000).

1.1 Prologue : d'un DESS de psycho à un Doctorat en Sciences du Langage

1.1.1 Un parcours

Mes années de lycée ont consolidé l'attrait qu'exerçaient pour moi les langues. Enfant, je voulais apprendre « plein de langues » : je me souviens avoir pioché dans une bibliothèque des dictionnaires bilingues et avoir démarré l'apprentissage de l'anglais et de l'espagnol par des exercices de mémorisation des mots que j'y lisais. Représentation naïve de l'enfance qui croit que maitriser une langue c'est en connaitre les mots ! Mais ces dictionnaires avaient l'intérêt de me mettre face à des unités tangibles de la langue, celles qu'un enfant de 8 ans peut précisément appréhender. Inutile de dire que l'exercice a tourné court, que la mémorisation n'a aucunement été efficace et que j'ai abandonné d'autant plus rapidement que je n'avais sous la main aucun locuteur sur lequel tester mes petites connaissances fraichement acquises. Mais cet épisode marquait le début d'un vif intérêt pour le langage.

On peut déduire sans peine de cet appétit juvénile pour les langues étrangères que mes réussites scolaires étaient plus du côté littéraire que du côté scientifique. Donc au moment de choisir une orientation au lycée, quoique mes envies auraient impliqué un cursus scientifique, la disparité de mes compétences dans les matières scientifiques et littéraires me dirigeait droit vers ces dernières. Et puisque j'aimais les langues, le choix d'un bac lettres-langues s'est naturellement imposé.

Cependant, à l'université, je n'avais pas envie de poursuivre par des études de langue, essentiellement parce que le seul débouché que je croyais envisageable à l'époque était l'enseignement et, à 18 ans, je ne souhaitais pas devenir enseignante – un refus que mes choix professionnels ultérieurs ont constamment démenti.

Mon premier cursus universitaire, à l'université Toulouse le Mirail (elle s'appelle maintenant Toulouse Jean Jaurès), m'a conduite à un DESS de psycho clinique, grâce auquel je fus embauchée dans une association d'éducation populaire, les CEMEA, pour y mettre en œuvre des dispositifs d'insertion pour des publics éloignés de l'emploi (le traitement social du chômage de la fin des années 80 et années 90). J'étais alors assez loin des langues et de la linguistique, mais parmi les actions auxquelles j'ai participé alors, un dispositif m'y a ramenée et m'a ramenée en même temps à l'université.

Nombre de demandeurs d'emploi « présentant des difficultés particulières d'insertion » fréquentant les stages que j'avais la responsabilité de mettre en œuvre étaient des migrants, maitrisant plus ou moins bien la langue orale et souvent très mal voire aucunement la langue écrite. Pour ce public spécifique, pour envisager une insertion professionnelle ou une formation quelconque, il était nécessaire d'accroitre significativement la maitrise de la langue, aussi bien orale qu'écrite. Là encore, je me suis rendu compte avec le recul que j'ai à l'époque fait preuve d'une certaine naïveté en m'improvisant formatrice dans ce domaine et en bricolant des séances d'apprentissage de la langue à partir de bouts de savoir glanés au CRDP[3] et lors de journées de formation organisées par la « Base Pédagogique de Soutien », une association proposant des ressources dans le champ de l'alphabétisation et de la lutte contre l'illettrisme.

Les stages dans lesquels j'intervenais mélangeaient en effet les deux types de publics : d'un côté, des migrants, voire plutôt des migrantes, pour la plupart arrivant de pays dans lesquels le taux d'alphabétisation était faible, en particulier pour les femmes – nombre d'endroits dans le monde considèrent qu'il n'y a pas lieu d'investir dans l'éducation des filles. Ceux-là n'étaient souvent pas alphabétisés dans leur propre langue, pas familiers d'une culture de l'écrit, et c'était tout ce travail d'entrée dans l'écrit qu'il fallait mener avec eux. Dans le même « fourre-tout » des politiques de lutte contre le chômage, d'un autre côté, était présent un autre public en difficulté avec l'écrit, mais bien différent car constitué de francophones ayant fréquenté l'école, et même l'école française, mais n'ayant pour autant pas appris à lire et à écrire, présentant donc ce que l'ANLCI (Agence Nationale de Lutte Contre l'Illettrisme) définit comme *illettrisme* :

> On parle d'illettrisme pour des personnes qui, après avoir été scolarisées en France, n'ont pas acquis une maitrise suffisante de la lecture, de l'écriture, du calcul, des compétences de base, pour être autonomes dans les situations simples de la vie courante. Il s'agit pour elles de réapprendre, de renouer avec la culture de l'écrit, avec les formations de base, dans le cadre de la politique de lutte contre l'illettrisme.[4]

L'illettrisme est classiquement distingué de *l'analphabétisme* par l'exposition à la scolarisation :

> On parle d'analphabétisme pour désigner des personnes qui n'ont jamais été scolarisées. Il s'agit pour elles d'entrer dans un premier niveau d'apprentissage.[4]

En principe, sur le plan pédagogique, les propositions de formation devraient être différenciées car les besoins des uns et des autres sont distincts. Mais, dans les faits, les stages se formaient de façon opportuniste avec les demandeurs d'emplois qui se manifestaient auprès de l'ANPE (ce n'était pas encore Pôle Emploi) en faisant état d'un besoin d'accroitre des compétences à l'écrit.

Je me suis donc à ce moment-là – le début des années 90 – confrontée à la délicate mission d'accompagner des publics adultes dans l'acquisition ou le réapprentissage de la langue écrite, sans formation particulière pour cette tâche. Dans le même temps, j'ai changé d'environnement et d'employeur, passant du milieu urbain au milieu rural. Ce n'est pas anecdotique, car les

3 Centre Régional de Documentation Pédagogique – Les CRDP ont été renommés en *Canopé*.

4 http://www.anlci.gouv.fr/Illettrisme/De-quoi-parle-t-on/Les-definitions page consultée le 6 avril 2017

personnes relevant de l'illettrisme plutôt que de l'analphabétisme se sont trouvées plus nombreuses en face de moi dans les ateliers que j'ai alors mis en place en réponse à une commande institutionnelle. Le travail que je menais dans ces ateliers est évoqué dans ma toute première publication, qui n'est à vrai dire pas une publication scientifique (Jacques, 1997), mais témoigne du souci de rendre compte de mon travail et de faire connaitre une expérimentation.

Au cours de ces ateliers, j'ai confirmé le sentiment que la bonne volonté et un effort d'autoformation n'étaient pas suffisants. La rencontre d'adultes ayant fréquenté l'école française sans parvenir à entrer dans la langue écrite s'imposa comme une énigme, une question qu'il me fallait débrouiller. Leur difficulté mettait en évidence ce que je n'avais pas perçu jusqu'alors : une différence fondamentale d'accès à la langue orale et à la langue écrite, l'une s'acquérant de façon naturelle par imprégnation – même si indéniablement la maitrise de l'oral n'est pas identique entre tous les êtres parlants –, l'autre étant si peu transparente qu'elle pouvait résister à l'apprentissage.

Pour élucider ces questions, j'ai pensé trouver des réponses dans un « Diplôme Universitaire de Formation à la Lutte contre l'Illettrisme » qui se mettait en place à l'université de Pau. Mais même si elle m'a permis de découvrir des contenus de qualité (notamment la sociologie de l'éducation et la sociolinguistique), cette formation m'a laissée sur ma faim théorique. Je n'avais toujours pas assez d'éléments pour comprendre ce que la langue écrite avait de spécifique qui la rendait si résistante.

Fin 1995, je reviens donc sur les bancs de l'université (toujours à Toulouse le Mirail) en Licence de Sciences du Langage, autorisée par dérogation à m'inscrire directement en Licence, sans devoir refaire un DEUG. J'y découvre toutes les facettes des Sciences du Langage, j'y acquiers la méthodologie de la recherche et m'y passionne pour des questionnements variés. Quoique salariée, je valide sans perdre de temps licence, maitrise et DEA, me constituant au passage un bagage informatique qui sera consolidé pendant ma thèse. Cette dernière débute en 1998, sous la direction d'Andrée Borillo, et conforte des directions de recherche diverses mais gouvernées par le double souci de cerner la langue attestée et d'élucider le fonctionnement en texte des unités étudiées, ce qui me conduisit assez vite à chercher à décrire certains aspects du fonctionnement des textes eux-mêmes.

Dès mes premiers travaux réalisés pour l'initiation à la recherche dans le cursus universitaire (mémoires de maitrise et de DEA, équivalents à M1 et M2), s'est ainsi dégagée une constante : travailler sur de la langue « authentique », dont il s'agit de comprendre et décrire les usages.

1.1.2 Une constante : corpus et données langagières attestées

Le parcours décrit précédemment avait en effet dirigé mon regard vers la langue écrite avec laquelle on traite, voire on bataille - quand on ne la maitrise pas -, au quotidien. Celle dont il m'intéresse de décortiquer les rouages, c'est la langue que l'on peut rencontrer au détour de sa vie personnelle ou professionnelle, pas la langue d'un sujet idéalisé, pas la langue travaillée du littérateur ou du poète.

Il faut reconnaitre dans ce choix la conjonction de ma curiosité, déjà évoquée, pour la langue et de l'empreinte sur moi du laboratoire de recherche dans lequel j'ai mené mes premiers travaux de recherche – l'Équipe de Recherche en Syntaxe et Sémantique (ERSS), unité mixte de recherche à l'Université Toulouse le Mirail –, et en particulier des chercheurs et enseignants-chercheurs que j'y ai rencontrés et auxquels j'exprime ma gratitude de la générosité intellectuelle qu'ils ont manifestée à mon égard. [Avoir Andrée Borillo pour directrice de thèse

est synonyme d'exigence mais surtout d'une formation riche et complète. Elle m'a « mise sur les rails » – si je peux me permettre cette expression – d'une description linguistique ancrée sur les données et a perçu très vite le potentiel de la linguistique outillée. C'est par son intermédiaire que j'ai eu l'occasion de travailler avec Anne Condamines, puis plus tard Didier Bourigault, tous deux alors fers de lance d'une terminologie textuelle, défendant l'idée que les discours sont les lieux d'élaboration des connaissances et de la langue d'une discipline, et que l'analyse terminologique doit en premier lieu se fonder sur l'analyse de corpus textuels (Bourigault & Slodzian, 1999). Anne Condamines a en outre toujours eu le souci d'un ancrage applicatif de la recherche – ce qui eut aussi le bénéfice non négligeable de générer des contrats de recherche, façon appréciable de concilier la nécessité de gagner sa vie quand on ne dispose pas d'allocation de thèse et une activité de recherche suivie. C'est à travers un partenariat qu'elle a mis en place avec des acteurs de la gestion des déplacements à Toulouse que je me suis impliquée dans une recherche bien éloignée de mes préoccupations initiales mais considérablement stimulante intellectuellement et qui a fourni la matière de ma thèse. Cette recherche m'a aussi permis de débuter avec Josette Rebeyrolle une collaboration fructueuse, autour d'abord de l'outillage linguistique car c'est elle qui m'a formée à l'utilisation du logiciel SATO[5], puis de la structuration textuelle, qui constitue une de mes principales directions de recherche. Josette Rebeyrolle collaborait elle-même avec Anne Condamines autour de l'élaboration de bases de connaissances terminologiques (Condamines & Rebeyrolle, 2000) et c'était une telle base qu'il s'agissait de bâtir pour les acteurs de la gestion des déplacements à Toulouse, en allant puiser dans leurs propres textes les termes et concepts pertinents.]

Il s'est donc très vite imposé comme une évidence que mon matériau serait formé de données langagières authentiques et attestées, collectées, « sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage » pour reprendre une définition qui a fait florès, traduite par Habert *et al.* (1997 : 11) de celle proposée par J. Sinclair, artisan de l'utilisation des corpus dans la linguistique britannique.

Se positionner comme « linguiste de corpus » est somme toute en 2017 assez trivial, l'évolution de la linguistique ayant rendu sinon caduque du moins exotique une approche introspective des faits de langue telle que la présentait par exemple Corbin (1980). Cependant, fin 99 début des années 2000, ce positionnement ne constituait pas encore la tendance majoritaire. Je développerai dans le chapitre qui suit les grandes lignes de ce positionnement comme linguiste de corpus et traiterai notamment la question de l'appréhension du corpus : source inductive de la description ou réservoir d'exemples ? Je tenterai d'argumenter en faveur d'une recherche pilotée par le corpus mais sans souscrire à l'illusion d'un usage a-théorique ou dénué d'à-priori. J'exposerai aussi la réflexion théorique et méthodologique que j'ai pu élaborer au fil de mes recherches, sur cette question de la production des données de l'analyse. Car ce n'est à mon sens pas le tout de constituer un corpus, encore faut-il être clair sur le statut des analyses que l'on y mène, donc être conscient de possibles biais méthodologiques.

Ma formation à l'ERSS s'est enrichie aussi de l'outillage logiciel et conceptuel qui permet tout à la fois de bâtir un corpus puis de le traiter (étiqueter et annoter) afin de l'exploiter. Grâce à Ludovic Tanguy et à l'équipe de « talistes », je me suis familiarisée avec le langage Perl (Tanguy & Hathout, 2007), et ai bénéficié d'un outil « clef en main » développé par Ludovic Tanguy, Yakwa (Rebeyrolle & Tanguy, 2001), pour mes investigations en corpus.

5 Il existe maintenant (juin 2017) une version en ligne de ce logiciel : http://www.ling.uqam.ca/sato/outils/sato.htm

Ces nouvelles cordes à mon arc m'ont conduite à des recherches à la croisée du TAL et de la linguistique, en collaboration avec deux laboratoires d'informatique, l'IRIT (Institut de Recherche en Informatique de Toulouse), qui est de longue date partenaire de l'ERSS pour de nombreux projets, puis le LIPN (Laboratoire d'Informatique de Paris Nord), à Villetaneuse, en banlieue parisienne, pour un contrat post-doctoral d'un an.

Les multiples compétences construites et les divers axes de recherche abordés autour d'une linguistique de corpus outillée ont été des atouts indéniables pour une intégration réussie au LIDILEM, où de nouvelles collaborations, avec Agnès Tutin, Laura Hartwell, Fanny Rinck, Catherine Brissaud et Francis Grossmann, en particulier, m'ont permis de poursuivre et approfondir certaines voies et d'en ouvrir de nouvelles.

Multiplicité des lieux, multiplicité des collaborations qui conduisent à envisager, à la faveur de rencontres intellectuellement stimulantes, de nouveaux objets de recherche, non sans lien avec les précédents… Les cheminements ont conduit à explorer plusieurs thèmes dont cette synthèse veut rendre compte.

1.1.3 Une ligne directrice : le texte et sa dynamique

Dans ce passage que cet écrit opère d'une cohérence subjective à une cohérence montrée, peu à peu une ligne directrice s'est dégagée qui devait organiser l'ensemble. J'ai choisi, comme on le verra à la fin de cette introduction, de découper cette synthèse en 5 chapitres, chacun couvrant une thématique de recherche. Quelle que soit la thématique, la ligne qui traverse mes travaux est d'ancrer la recherche non seulement sur corpus mais en prenant en compte la dynamique du texte. Il n'en ressort pas de propositions théoriques quant au texte et à la linguistique textuelle, mais des travaux qui ne considèrent pas les textes des corpus sur lesquels j'ai travaillé comme des « sacs de mots » (position encore trop souvent répandue dans le Traitement Automatique des Langues) ou comme des patchworks de micro-textes, tissés de relations de phrase à phrase (comme trop de théories formelles de la cohérence textuelle le laissent penser), donc des travaux qui « prennent au sérieux » la dimension textuelle, en considérant que le texte est un support matériel comportant des instructions pour son traitement cognitif (section 1.2).

On verra que, selon les thématiques, « prendre le texte au sérieux » a signifié observer dans des textes spécialisés les modifications formelles des termes complexes du domaine, analyser le rôle de certains objets textuels, décrire la maitrise de certains procédés rhétoriques par des étudiants francophones et allophones, participer à la construction d'un projet de recherche visant – entre autres – la description du développement des compétences de production d'écrit tout au long de la scolarité… Quelle que soit la thématique, le texte en tant qu'unité donne cadre et sens aux phénomènes observés, qu'ils soient appréhendés à un niveau « micro » – d'une phrase à l'autre –, à un niveau « macro » – l'organisation globale du texte – ou à un niveau « supra » – l'univers de connaissances spécialisées dans lequel le texte s'inscrit.

Le titre choisi pour cette synthèse me place dans une conception résolument dynamique qui n'épuise cependant pas la notion de textualité – à vrai dire, en rédigeant ceci, je prends conscience de ce que mon travail s'ancre dans une certaine notion de la textualité que je n'ai certainement pas la prétention de théoriser, car elle n'est pas suffisamment poussée et élaborée pour constituer un système cohérent apte à rendre compte de toutes les facettes de cet objet complexe qu'est le texte. Sans doute un regard extérieur y percevra des zones d'ombre, des impensés qui tiennent à ce que sont laissées de côté des questions et des problématiques qui ne sont pas directement reliées aux objets que je travaille. Si je voulais, comme le font par exemple

depuis de nombreuses années J.-M. Adam ou F. Rastier dans le contexte francophone, Van Dijk dans le contexte anglophone, proposer une théorie de la textualité, toutes les facettes de l'objet devraient être explorées, ses propriétés et caractéristiques élucidées.

On le verra à la lecture de cette synthèse, le travail que je présente ne vise pas à faire progresser la compréhension de ce qu'est le texte en tant que tel mais le prend comme objet empirique, inscrit dans une matérialité porteuse de sens (cf. chapitre 4) et support d'opérations mentales. S'en distingue la notion de discours, ce qui mérite explicitation.

1.2 Cadre théorique : texte et discours

Nombre de linguistes et de théories linguistiques proposent une distinction entre le texte et le discours. Mais les définitions du texte et du discours, et par voie de conséquence la différenciation, voire l'opposition entre ces deux notions, sont hautement variables. C'est donc une question pour laquelle il faut d'abord mettre un peu d'ordre afin d'y voir plus clair et assoir un positionnement intelligible et cohérent.

Le terme de *discours*, dans le contexte de la linguistique française, vient volontiers en association avec *analyse*, pour former *analyse de discours*. D'après (Dufour & Rosier, 2012), l'analyse de discours « à la française » (ADF) ne s'est pas en premier lieu fondée sur une distinction avec une linguistique du texte – qui dans les années 1960 n'était pas non plus constituée en tant que *linguistique textuelle*, telle que celle-ci se présente en ce début de 21e siècle – mais a œuvré à se définir en propre, au carrefour de différentes disciplines :

> L'une des caractéristiques de l'ADF, témoignant à la fois de sa plasticité conceptuelle et de ses mouvements théoriques, est sa façon régulière de développer une réflexion épistémologique sur la constitution de son champ d'analyse, sur les ancrages et outils mobilisés à un moment donné dans un contexte donné, de même que sur sa visée historiquement émancipatrice et plutôt orientée aujourd'hui à la satisfaction d'une demande sociale. (Dufour & Rosier, 2012 : 6)

Les deux auteures mettent en avant le problème du champ d'analyse, la réflexion sur les outils et les méthodes, l'intégration des filiations et références théoriques comme cadre pour les analyses, et on perçoit en filigrane que le travail sur la différenciation et les relations des deux notions *texte* et *discours* n'était pas pertinent pour la fondation de l'analyse de discours. C'est le discours d'institutions, dans son contexte social, qui est revendiqué par l'ADF comme champ d'études :

> cette étiquette [l'école française d'analyse du discours] permet cependant de circonscrire un champ d'études à l'origine essentiellement portées sur des écrits émanant d'institutions officielles (discours politique, syndical, puis médiatique), marquées par le marxisme et la psychanalyse. Elle est emblématique d'une conception du discours traversé par d'autres discours qui le constituent en se constituant, et d'un sujet dominé, traversé lui aussi par une altérité irréductible. (Dufour & Rosier, 2012 : 3)

Le discours tel que travaillé par l'ADF serait donc le lieu pour l'étude des marques de l'énonciation et de l'inscription dans une pratique sociale.

> L'ADF, en particulier dans sa « seconde mouture », s'inscrit dans une veine énonciative, héritée de Bally, Benveniste, Maingueneau, Rabatel alors que l'analyse de discours anglo-saxonne est davantage ancrée dans la pragmatique. (Dufour & Rosier, 2012 : 11)

La distinction *texte/discours* qui se dessine semble ainsi assez classique : le discours conçu comme « le texte (ou l'énoncé, comme on va le voir ci-dessous) plus le contexte énonciatif » et, à l'inverse, le texte correspond au « discours moins le contexte énonciatif ». C'est en ces termes que F. Rastier rend compte d'une différenciation qu'en même temps il juge non nécessaire :

> Le texte n'est qu'une sorte d'énoncé. Charaudeau (1973[6], p. 28) synthétisait déjà ainsi l'opposition énoncé/discours.
>
> ÉNONCÉ + situation de communication = DISCOURS
>
> usage - consensus sens
>
> spécificité signification (Rastier, 2005)

Pour Rastier, la séparation *texte/discours* est affaire de politique académique et de constitution de champ :

> La sémiotique discursive a postulé cette séparation pour se créer un objet disciplinaire indépendant de la linguistique. Il s'agit là toutefois de politique académique plus que de problématique scientifique (Rastier, 2005)

Je lui laisse la responsabilité de ce propos tout en notant qu'en effet, vers la fin des années 60, le dépassement du cadre devenu trop étroit de la phrase et l'intégration de l'objet *discours* comme objet de la linguistique devait s'accompagner d'une construction intellectuelle des spécificités de cet objet et de la création d'un territoire scientifique propre.

C'est ce que fait le numéro fondateur que la revue *Langages* consacre en 1969 à « L'analyse du discours ». Les articles y bâtissent et défendent un projet scientifique et une méthodologie visant à élargir le champ de la linguistique au-delà de l'horizon de la phrase, et ce en mettant en évidence des règles d'enchainements des énoncés au sein du texte.

> 1.2. La séquence des phrases constitue l'énoncé qui devient discours lorsque l'on peut formuler des règles d'enchaînement des suites de phrases. Certes, si l'on borne la linguistique à la phrase, on renverra les règles du discours à d'autres modèles et à d'autres méthodes, en particulier au sujet parlant et à la psychologie; mais si l'on considère que la linguistique doit rendre compte des énoncés produits et/ ou productibles, il convient de définir des enchaînements selon les méthodes et les principes qui ont présidé à l'établissement des règles de la phrase. (Sumpf & Dubois, 1969 : 3)

Dans l'article qui ouvre le numéro, Z. Harris met assez naturellement en équivalence « énoncé suivi (écrit ou oral) » et « discours » (Harris, 1969 : 8) et poursuit en utilisant le terme *texte* sans toutefois expliciter le rapport que le texte entretient avec le discours. Il expose la méthode (la construction de classes d'équivalences des énoncés et leur analyse distributionnelle) qui permet précisément d'atteindre l'objectif évoqué par Sumpf et Dubois : la formulation de règles d'enchainement des suites de phrases.

Il est intéressant de souligner que c'est une ambition similaire que poursuivent les travaux sur les grammaires textuelles, assez mal nommées, expliquent Charolles et Combettes (1999), car elles ne visent pas l'établissement d'un système de règles de bonne formation qui s'appliqueraient au texte entier comme il en existe pour la phrase – en ce sens elles ne sont pas une *grammaire* – mais elles ambitionnent de formuler les contraintes qui régissent

6 Il semblerait qu'il y ait ici une coquille et qu'il s'agisse plutôt de Charaudeau, P. (1983) *Langage et discours*, Paris, Hachette.

l'organisation des phrases et en font un *texte*. L'hypothèse sous-jacente est qu'il existe chez le sujet parlant une compétence textuelle comme il existe une compétence phrastique, et

> [d]e là découle l'idée d'appeler *« grammaire de texte »* tout modèle ayant pour vocation, dans le prolongement de la grammaire de phrase, d'expliciter les règles sur lesquelles s'appuient les sujets parlants pour distinguer les suites de phrases (acceptables comme phrases) mais sans rapport entre elles et celles qui forment un discours suivi. (Charolles & Combettes, 1999 : 83-84)

Avec l'analyse du discours et les grammaires de texte, se dessinent deux cadres d'analyse qui emploient tous deux les termes de discours et de texte. D'un côté, l'analyse de la relation aux institutions et à la société met l'accent sur le rapport de l'énoncé à ses conditions de production ; d'un autre côté, l'étude de la façon dont les énoncés / propositions (les deux termes apparaissent, employés par l'un ou l'autre) forment ensemble un texte dirige le regard vers les phénomènes de cohésion textuelle (par exemple Halliday & Hasan, 1976), de suivi et de progression thématique (Combettes, 1988), et donc vers la structuration textuelle à tous les niveaux (Van Dijk, 1995). Mais pour Maingueneau, cette dichotomie ne « tient » pas dans la mesure où le discours noue le texte et un lieu social :

> L'intérêt qui gouverne l'analyse du discours, ce serait d'appréhender le discours comme intrication d'un texte et d'un lieu social, c'est-à-dire que son objet n'est ni l'organisation textuelle ni la situation de communication, mais ce qui les noue à travers un dispositif d'énonciation spécifique. Ce dispositif relève à la fois du verbal et de l'institutionnel : penser les lieux indépendamment des paroles qu'ils autorisent, ou penser les paroles indépendamment des lieux dont elles sont partie prenante, ce serait rester en deçà des exigences qui fondent l'analyse du discours. (Maingueneau, 2005 : 66)

Dans sa présentation à ce numéro de *Marges Linguistiques* destiné à proposer un état de l'art de l'analyse du discours – environ 40 ans après le numéro de *Langages* qui porte ce titre –, Maingueneau distingue une linguistique du texte, celle-là même qui ambitionne de rendre compte de la *textualité*, et une linguistique du discours, dont l'analyse du discours n'est qu'une des composantes. En filigrane se recomposent les deux notions : le texte comme enchainement suivi et cohérent de propositions, doté d'une organisation, le discours comme dispositif d'énonciation spécifique.

Suspendant la question de l'énonciation et de l'inscription dans un « lieu » (au sens de Maingueneau ci-dessus), une linguistique *textuelle* s'est développée, soucieuse de la description des propriétés des textes. Elle met au rang de ses interrogations une mise en ordre classificatoire, de nombreuses typologies mettent l'accent sur telle ou telle caractéristique (Petitjean, 1989). Il s'agit de cerner des types et leurs modes d'organisation spécifiques (Adam, 1992, 2004) en mettant au jour les corrélats linguistiques de ces types.

J'ai pour ma part suivi une direction différente, qui, je l'ai dit, met au premier plan la dimension dynamique du texte vu non comme un donné mais comme support d'une construction, ce qui me conduit à employer *discours* dans un sens autre, sans relation avec l'analyse du discours, je l'explicite ci-après.

Charolles et Combettes (1999), quoique le titre de leur article évoque « l'analyse du discours », montrent comment une linguistique du texte a progressivement glissé de l'analyse des facteurs de textualité que sont les indices de cohésion textuelle (anaphores, connecteurs, continuités et ruptures thématiques, etc.) vers une appréhension plus cognitive du texte et des effets de ces marques sur la compréhension. T. Van Dijk, par exemple, travaille avec le psychologue

W. Kintsch pour élaborer sur une assise cognitive un modèle du traitement des textes par le lecteur. Van Dijk et Kintsch font l'hypothèse que le texte donne lieu à la construction par le lecteur d'une macro-structure, base de la compréhension mobilisée dans des tâches de rappels. Un de leurs premiers articles (écrit en français) pose une définition du *texte* sans préciser toutefois ce qui est entendu par *discours* (c'est moi qui souligne les deux termes dans les extraits qui suivent).

> On appellera **texte** la structure formelle, grammaticale, d'un **discours**. (Kintsch & Van Dijk, 1975 : 100)

> Un **texte**, de façon intuitive, ne se réduit pas à la séquence des phrases qui le constituent. De la même façon qu'une phrase est interprétée et traitée plus avant en fonction des structures hiérarchiques sous-jacentes, de même un **discours** est interprété, stocké, rappelé en fonction de sa structure d'ensemble que nous appelons sa *macro-structure*. (Kintsch & Van Dijk, 1975 : 101)

La question est là déplacée vers la question de l'interprétation. Ce qui guide la compréhension et la mémorisation d'un texte, c'est que l'on peut en bâtir une interprétation qui est fonction d'une composante interne du discours : sa structure d'ensemble ou macro-structure, qui donne prise sur le discours en ce qu'elle indique au lecteur les macro-propositions qui composent le texte et lui permettent de faire le départ entre l'essentiel et l'accessoire – l'accessoire étant ce qui n'a pas d'incidence sur le schéma général de l'histoire (si le texte est une narration) ou de l'argumentation (si la visée du texte est argumentative). L'évolution de leur théorie les conduit à envisager une part plus grande pour l'activité du lecteur. Celui-ci n'est pas le réceptacle d'une interprétation programmée entièrement par les structures du texte, il construit un **modèle mental** du texte en mobilisant aussi ses connaissances du monde, ses affects, etc.

Cette version « cognitive » qui distingue le texte comme succession d'énoncés « donnée » au lecteur pour que celui-ci en construise une représentation adéquate, i.e. un modèle mental, est précisément celle sur laquelle sont bâtis de nombreux travaux de recherche linguistique et psycho-linguistique, comme on va le voir.

Elle conceptualise le texte comme le substrat matériel porteur aussi bien d'un contenu sémantique et pragmatique – que Kintsch et Van Dijk ont dans un premier temps conçu comme la base sous-jacente du texte, *textbase* (Van Dijk, 1995) – que d'« instructions » destinées au traitement cognitif du texte. Un parfait exemple de cette dimension instructionnelle du texte est fourni par l'analyse des anaphores et de leur résolution. L'anaphore (linguistique) peut être conçue comme la simple mise en relation entre une expression anaphorique référentiellement incomplète, telle qu'un pronom personnel comme *il* ou *elle*, et une expression textuelle qui la précède, par exemple :

(1) J'ai vu le dernier film primé à Cannes. **Il** est bouleversant. [exemple fabriqué]

Ici, *il* renvoie à *le dernier film primé à Cannes*. Or de nombreux cas d'anaphores ne font pas intervenir un antécédent textuel explicite, mais mettent en jeu un référent implicite, supposément accessible parce que le lecteur (ou l'auditeur en cas de communication orale) est à même de l'introduire dans son modèle mental pour résoudre l'anaphore. Imaginons l'échange fictif suivant :

(2) - L'aide au logement diminue de 5 euros !
- Oh mais, vous savez, **ils** ne se rendent pas compte, ce que c'est 5 euros, quand on gagne le smic. [exemple fabriqué]

Ce *ils* qui renvoie indistinctement aux politiques, à ceux qui font les lois et prennent les décisions, est un exemple typique qui met à mal la notion d'antécédent textuel (Kleiber, 1990b) et montre qu'il y a dans le modèle mental bien plus que ce que le texte apporte. Un tel exemple plaide pour une conception de l'anaphore comme déclencheur d'une opération mentale de construction d'un objet-de-discours (Apothéloz, 1995a) à même de fournir le référent approprié pour l'expression à résoudre.

La défense de cette thèse de l'anaphore déclencheur d'une opération de construction est le cheval de bataille de F. Cornish (1990, 1996, 2001, 2003, 2010). Ses travaux l'ont peu à peu conduit à préciser les relations entre le texte, le contexte et le discours, je reproduis ci-dessous un tableau synthétisant ces trois notions[7].

Text	Context	Discourse
The connected sequence of verbal signs and non-verbal signals in terms of which *discourse* is co-constructed by the discourse partners in the act of communication.	The *context* (the domain of reference of a given text, the co-text, the discourse already constructed upstream, the genre of speech event in progress, the socio-cultural environment assumed by the text, and the specific utterance situation at hand) is subject to a continuous process of construction and revision as the discourse unfolds. It is by invoking an appropriate context that the addressee or reader may create *discourse* on the basis of the connected sequence of textual cues that is *text*.	The product of the hierarchical, situated sequence of utterance, indexical, propositional and illocutionary acts carried out in pursuit of some communicative goal, and integrated within a given *context*.

Table 1. The respective roles of *text*, *context* and *discourse* (Cornish, 2008:Table 1, p. 998, revised)

Tableau 1 : *Tableau synthétisant les notions de texte, contexte, discours* (Cornish, 2010)

Cette synthèse montre clairement la distinction entre le *texte*, conçu comme le substrat matériel, et le *discours*, comme le produit résultant de son traitement, celui-ci intégré au *contexte*. Pour Cornish (2010), le texte est une séquence de signes verbaux (oraux ou écrits) et représente la trace matérielle d'un acte d'énonciation, ce qui signifie que les signes qui le composent sont à prendre autant comme véhicules du sens conventionnel qui est attaché aux unités (lexicales, grammaticales) que comme indices de la signification à construire. Celle-ci n'est pas donnée, elle est donc le résultat d'une construction, qui fait appel au contexte pertinent, qui lui-même se compose de multiples dimensions telles que le co-texte, le domaine de référence (les connaissances d'arrière-plan), le genre, la situation, etc. L'exploitation du texte et du contexte construit le discours : un modèle mental dynamique, en perpétuel remaniement au fur et à mesure de l'avancée de la communication, écrite ou orale, dans lequel tout n'est pas sur le même plan, mais structuré et hiérarchisé.

> Discourse, then, is both hierarchical and defeasible (a provisional, and hence revisable, construction of a situated interpretation), whereas text is essentially linear [...]. Discourse clearly depends both on *text* **and** *context* . It is the discourse constructed in terms of the text and a relevant context which is capable of being

7 Je n'ai malheureusement eu accès à cet article que dans sa version électronique et ne puis indiquer les numéros de pages précis des citations. L'article est en ligne à cette adresse : https://halshs.archives-ouvertes.fr/hal-00966398/document (consulté le 01/09/2017).

> stored subsequently in long-term memory for possible retrieval at some later point. The textual trace of the communicative event, for its part, is short-lived, disappearing from short-term memory once that discourse is constructed — or very soon thereafter (cf. Jarvella, 1979). See also Ariel (2008: 2), Langacker (1996: 334) and Widdowson (2004: 8). Text, context and discourse, then, are interdependent, interactive and inter-defining. (Cornish, 2010)

J'adopte tout à fait cette conception cognitive de la fabrication d'une représentation hiérarchisée, par le lecteur puisque je centre mes travaux sur l'écrit, représentation appuyée sur les indices présents dans le texte – notamment le sens instructionnel de certaines expressions linguistiques – et sur les éléments pertinents du contexte.

Pour dénommer ce modèle mental, qui implique plus étroitement le lecteur et sa perception du texte et du contexte, le terme de *discours* a été choisi par F. Cornish. Je le retiendrai à sa suite, en étant consciente de la possible ambigüité avec le sens de *discours* porté par le terme *analyse de discours* et de la difficulté qu'il y a à maintenir tout au long d'un écrit une séparation nette entre le texte comme substrat matériel et le discours comme construit, dans la mesure où le discours n'est pas seulement du côté du récepteur – comme pourraient le laisser croire les formulations précédentes, mais est présent comme virtualité dans le texte même. En atteste l'existence d'unités linguistiques dont le sens est tout autant instructionnel que descriptif (Kleiber, 1994). Par exemple, les déterminants donnent des indications sur la façon de construire la référence : un emploi de l'article défini (*le, la, les*) signale une certaine unicité du référent (de Mulder, 1994) qui guide ainsi cette construction.

De même, les travaux sur la notion d'accessibilité reposent tout entiers sur la conception cognitive du traitement du texte (de Mulder, 2000 ; Gundel, 1998 ; Gundel, Hedberg & Zacharski, 2000 ; Walker, 1998). La conception sous-jacente à la notion d'accessibilité est que le choix des expressions linguistiques est fonction des hypothèses que fait l'auteur d'un texte (que celui-ci soit écrit ou oral) sur le modèle mental du récepteur du texte à un moment donné. Si je prononce au petit déjeuner la phrase suivante :

(3) Tu as donné des croquettes au chat ? [exemple pas si inventé que ça]

c'est parce que je fais l'hypothèse que mon interlocuteur pourra sans peine « récupérer » le référent, soit que le chat en question est présent dans la pièce, soit que le fait que nous n'ayons qu'un chat le rend définitivement hautement accessible. Si mon interlocuteur me répond « Quel chat ? », c'est raté, ce qui prouve que l'accessibilité est bien une affaire de présupposition de la part de celui qui élabore le texte et non un donné indéfectible du texte ou de la situation[8].

Une part de mes recherches prend explicitement pied sur le caractère instructionnel du texte et tente d'en inférer un discours possible.

La position que je viens d'exposer ne signifie un désintérêt ni pour les propositions de la linguistique textuelle quand elle réfléchit aux propriétés des textes – on verra dans le chapitre 5 que la question de la relation du texte au genre a sa place dans mon travail –, ni pour

8 Il a souvent été reproché aux théories de l'accessibilité une forme de circularité : c'est parce qu'un référent est encodé avec une expression qui marque une haute accessibilité qu'il est très accessible et c'est parce qu'il est très accessible qu'il est encodé avec une expression qui marque une haute accessibilité. Or la conversation quotidienne regorge d'ajustements mutuels pour l'identification des référents sous la forme de demandes de précisions de qui sont les *il* ou les *elle*, qui montrent que l'expression choisie par le locuteur n'est aucunement un gage de la position effective du référent sur l'échelle d'accessibilité, mais correspond à la supposition qu'il fait de cette position. La même situation peut se reproduire à l'écrit, mais comme la communication est *in absentia*, le réajustement n'est pas possible.

l'approche de l'ADF. Pour les objets et phénomènes que j'ai travaillés, en particulier la réduction des termes complexes (voir chapitre 3) et le rôle des intertitres (voir chapitre 4), la perspective la plus féconde m'a semblé être celle qui repose sur un modèle de dynamique textuelle mettant en jeu une dimension cognitive, tel que je viens de l'exposer.

1.3 Directions de recherche

J'ai indiqué dans le titre des directions de recherche mais il s'agit plutôt de trois niveaux pour des recherches qui s'articulent et se complètent et, pour finir, trouvent un prolongement dans des applications :

1er niveau : cadrage méthodologique → réflexion des choix méthodologiques, qui me semble incontournable pour assurer la validité des descriptions menées ;

2e niveau : le lexique → étude au sein du texte spécialisé de certaines des unités lexicales qu'il mobilise, les termes et le lexique scientifique transdisciplinaire ;

3e niveau : la structuration du texte → description des fonctions de certains objets textuels, les titres de section.

Une première partie de ce mémoire d'habilitation rendra explicite et ces trois niveaux et les résultats obtenus, une seconde partie envisagera la dimension applicative. Les chapitres s'organisent donc comme suit.

1.3.1 Méthodologie et descriptions

1. Réflexions méthodologiques

J'y ai déjà fait allusion, l'essor dans les sciences du langage, en particulier ces quinze dernières années, d'une production de données appuyée sur des corpus textuels ne me semble pas devoir aller sans questionnement sur le statut des données, car ce n'est pas parce que les données proviennent de sources authentiques qu'elles doivent être prises en considération sans précaution. Le **chapitre 2** sera consacré à la clarification des références à partir desquelles j'aborde le travail en corpus, essentiellement les travaux britanniques en linguistique de corpus (Chafe, 1992 ; Fillmore, 1992 ; Habert *et al.*, 1997 ; Halliday, 1992 ; Leech, 1992 ; Léon, 2008), et aux précautions méthodologiques qui me semblent absolument nécessaires pour la validité d'une linguistique de corpus.

Cet appareil théorique et méthodologique cadre l'ensemble de mes recherches, des unités lexicales à l'unité texte.

2. Études du lexique

Il y a mille et une façons d'étudier le lexique et ses unités. On peut réfléchir à une typologie, un classement, aux propriétés d'une catégorie grammaticale (Borillo, 1997 ; Flaux & Van de Velde, 2000 ; G. Gross, 1994), ou de façon plus générale (Le Pesant & Mathieu-Colas, 1998) ; on peut s'interroger sur le sens : existe-t-il, n'existe-t-il pas, comment se construit-il, comment se décrit-il (Cadiot, 1994 ; Cruse, 1986 ; Kleiber, 1997a, 1997c ; Rastier, 1994 ; Récanati, 2001) ? On peut théoriser sur la polysémie, l'ambigüité et le rôle du contexte (Charolles, 1996 ; Fuchs, 1986, 1991, 1997, Kleiber, 1994, 1997b, 1999 ; Victorri & Fuchs, 1996), ou sur la néologie et l'innovation lexicale (Bastuji, 1974 ; Cusin-Berche, 1999 ; Guilbert, 1975 ; Sablayrolles, 2000), ou selon d'autres perspectives encore, qui vont de l'analyse des unités

elles-mêmes à l'examen de leurs conditions d'emploi en passant par la réflexion sur la façon la plus appropriée de les décrire et de les définir dans des dictionnaires ou des glossaires...

La lexicologie est un vaste territoire dont je n'ai travaillé qu'une micro parcelle en circonscrivant très précisément l'univers considéré, celui de la langue spécialisée, et les unités regardées, i. les termes, unités porteuses de la connaissance d'un domaine spécialisé, ii. les unités du lexique scientifique transdisciplinaire, le lexique qui permet de « parler science » (comme « parler français » ou « parler espagnol »). Le **chapitre 3** explicitera l'angle d'analyse sous lequel les termes ont été regardés : leur comportement dans les textes. C'est toujours à partir des textes qu'a été travaillé le lexique scientifique transdisciplinaire (LST), car ce sont les contextes qui témoignent de l'appartenance d'une unité à la catégorie du LST. Ce chapitre est une première illustration de la permanence, dans l'ensemble de mes recherches, de l'intérêt pour les usages spécialisés de la langue, l'écriture scientifique étant l'un de ces usages. Il montre aussi la mise en pratique de l'approche méthodologique défendue au chapitre 2 et qui cadre les recherches sur la structuration du texte.

3. La structuration du texte

De quels moyens dispose un scripteur pour mettre en évidence une organisation, un découpage, une hiérarchisation, qui font qu'un texte ne se lit et ne se comprend pas seulement en cheminant au long de son déroulement mais en cernant sa progression, en saisissant ses relations internes, ses parallèles, ses décrochages, ses emboitements ? Diverses études portent sur les moyens verbaux tels que les marqueurs d'intégration linéaire, qui permettent de construire des séries, d'établir des parallèles, d'organiser la successivité (Jackiewicz, 2004), ou les cadres de discours qui délimitent et organisent des séquences discursives (Charolles, 1997 ; Jackiewicz & Minel, 2003). Mais dans des textes très structurés, articles de recherche, mémoires, thèses, documents professionnels tels que projets et compte-rendus, la macro-structure, qui concerne le texte dans son ensemble, est partiellement livrée par un moyen visuel : le découpage en sections et sous-sections, qui, non content de segmenter le texte, l'augmente par une opération de titrage de ces sections et sous-sections. Ces titres de sections, que nous avons dénommés *intertitres*, ne font l'objet que de très peu de travaux. Le **chapitre 4** comblera cette lacune en explicitant les points de départ théoriques et les échafaudages méthodologiques qui ont été mis sur pied pour l'étude des intertitres. L'ambition était de bâtir un cadre qui permette d'appréhender les fonctions des intertitres dans les textes. Ces objets spécifiques du texte écrit sont en effet reconnus par les psychologues comme ayant un impact sur le traitement cognitif du texte, mais il revient aux études linguistiques de cerner plus finement leur(s) rôle(s). L'appareillage descriptif que nous avons élaboré[9] se fonde délibérément sur des traits formels. Une des raisons en est que la description se veut exploitable pour d'éventuelles tâches de traitement automatique. C'est là une autre constante de mes travaux : ils envisagent de donner lieu à des applications.

1.3.2 Pour une recherche guidée par les applications

Une part des études menées selon ces directions est en effet justifiable de l'appellation *linguistique appliquée* car elles trouvent un prolongement vers des réalisations concrètes sur les deux terrains qui ont, d'après (Léon, 2015), historiquement constitué la linguistique appliquée en France : le traitement automatique des langues et l'enseignement. De l'un à l'autre, mon travail ne procède pas par rupture, comme on le verra dans le **chapitre 5**. Partant

9 Le « nous » renvoie ici au petit groupe qui s'est emparé de cet objet d'étude : Lydia-Mai Ho-Dac, Josette Rebeyrolle, Marie-Paule Péry-Woodley et moi-même.

d'une linguistique outillée et de problématiques de traitement automatique, l'analyse syntaxique automatique puis le repérage en contexte de passages exprimant des relations lexicales, j'ai transféré les démarches et outils élaborés pour ces traitements vers un questionnement plus didactique. Utilisant les techniques de TAL et l'appui sur des corpus, c'est d'abord une ressource pour faciliter l'accès à des passages précis d'articles scientifiques en anglais de la base Scientext qui est proposée.

Le **chapitre 6** amplifie cette orientation didactique et la prolonge vers l'avenir en exposant les motivations pour la réalisation d'un corpus d'écrits universitaires, Littéracie Avancée, destiné à être l'un des composants d'un grand corpus d'écrits scolaires, « de la maternelle à l'université », dont la constitution aura pour cadre un projet financé par l'ANR[10]. Ce dernier chapitre a donc une dimension très prospective car il s'appuie sur les réalisations déjà achevées pour esquisser les réalisations futures. Il met en lumière un positionnement qui m'est cher : un retour « vers la société » qui tire profit des avancées d'une discipline comme la linguistique. Et d'une certaine manière, ce chapitre « boucle la boucle » : ainsi que je l'écrivais dans la section 1.1, je suis venue à la linguistique à l'occasion d'une interrogation sur la résistance intrinsèque de la langue écrite pour des apprenants. Je reviens à l'enseignement de la langue et de l'écriture avec des interrogations différentes mais une même conviction que la recherche, quand elle le peut, doit savoir apporter des réponses pour de tels enjeux sociétaux.

Mes travaux se sont fondés sur divers corpus, constitués en fonction des objets de recherche à travailler, soit par mes soins, soit au sein d'une équipe de chercheurs. Pour l'intelligibilité du document, la section qui suit répertorie de façon synthétique ces divers corpus et renvoie au chapitre dans lequel chacun est décrit et justifié. On pourra aussi se reporter à la page 9 pour une liste synthétique.

1.4 Les corpus travaillés

Mes recherches ne s'appuyant pas sur des mesures lexicométriques, je n'indiquerai pas systématiquement ici la taille des corpus en nombre de mots, donnée peu pertinente pour les objets que j'analyse, mais donnerai plutôt le nombre de textes du corpus, avec une indication de la longueur des textes en nombre de pages – équivalent de page A4.

1. [Déplacements] et [Vol Libre]

Le corpus [Déplacements] (recueilli par A. Condamines et présenté page 54) était constitué de textes élaborés dans un contexte professionnel par quatre acteurs de la gestion des déplacements sur l'agglomération toulousaine : Mairie de Toulouse, DDE (Direction Départementale de l'Équipement), SEMVAT (compagnie de transports en commun), SMTC (Syndicat Mixte de Transports en Commun). Il rassemble par exemple le texte du projet de gestion globale des déplacements qui réunit ces quatre acteurs : des règlements administratifs pour des équipements à réaliser dans le cadre de ce projet, découpés en articles, des descriptions de systèmes informatiques impliqués dans cette gestion, etc. Les textes sont assez longs : une dizaine à une soixantaine de pages. Le corpus était confidentiel – en raison de règlements d'appels d'offre – et je n'en ai conservé qu'une partie : les documents de la DDE et ceux de la Mairie de Toulouse. Quand j'évoque le corpus [Déplacements], ce sont ces deux derniers ensembles de textes qui sont concernés.

10 Agence Nationale pour la Recherche, agence étatique qui assure le financement de la recherche sur projets.

Le corpus [Vol Libre] (présenté page 56) est constitué d'une vingtaine d'articles parus dans la presse spécialisée du vol libre, les deux magazines *Parapente Mag* et *Vol Libre*, dans la rubrique « essais de voiles ». Chaque article est un texte de deux à quatre pages.

Les deux corpus ont été fournis au format électronique et ont subi un étiquetage morphosyntaxique avec TreeTagger (Schmid, 1994). Ils ont servi de support aux analyses sur la réduction des termes complexes (section 3.1.2.2). Le sous-corpus issu de la DDE a aussi été utilisé pour les premières études sur les intertitres (section 4.2).

2. [Bricolage] et [Médecine]

Les deux corpus ont été constitués par aspiration de pages web, par l'un des partenaires du projet ANR TextCoop, auquel j'ai participé au sein du LIPN (Laboratoire d'Informatique de Paris-Nord), et qui visait à explorer les textes procéduraux.

Le corpus [Médecine] (page 45) est constitué d'environ 80 recommandations médicales, qui sont des textes élaborés par des médecins chercheurs à destination des praticiens en vue de diffuser auprès de ces derniers les avancées de la recherche médicale et de leur permettre d'adopter de bonnes pratiques médicales. La taille des textes varie de 5-6 pages à une vingtaine.

Le corpus [Bricolage] (page 45) est constitué de fiches pratiques mises à disposition par les sites web Mr Bricolage et Maison Facile, pour aider le lecteur aussi bien pour des réalisations techniques (comme poser un carrelage ou repeindre un radiateur) que pour des actes du quotidien (comme régler un litige avec un voisin). Chaque fiche équivaut à une à deux pages de texte.

Les pages ont été récupérées au format électronique et nettoyées de tout élément périphérique tel que publicité pour le site ou autre.

Ces corpus ont été utilisés pour l'étude de la formulation des procédures, laquelle a impliqué la définition d'une méthodologie adaptée (voir section 2.2).

3. [articles scientifiques]

Il s'agit en fait de deux corpus d'articles scientifiques (voir page 86), l'un dans le domaine de l'Ingénierie des Connaissances, l'autre dans le domaine de la Géopolitique, qui ont été constitués respectivement par J. Rebeyrolle et par L.-M. Ho-Dac pour leur propre thèse, et qui ont été versés dans un « pot commun » avec le corpus Déplacements pour les travaux que nous avons menés ensemble sur les intertitres (chapitre 4).

Les articles d'Ingénierie des Connaissances sont issus de la conférence Ingénierie des Connaissances, qui a lieu tous les ans, et respectent le format « classique » d'un article scientifique : une vingtaine de pages.

Les articles de Géopolitique ont été aspirés sur le site web de l'IFRI[11] (Institut Français des Relations Internationales) et sont un peu plus longs : de 20 à parfois 30 pages.

La nature des textes, articles scientifiques, et leur longueur, au moins une vingtaine de pages, ont pour conséquence une structuration en sections et sous-sections qui a été notre échelle de mesure : nous avons prélevé dans chacun des trois corpus le nombre de textes nécessaires pour atteindre un nombre suffisant et équivalent d'intertitres : environ 350 pour chaque sous-corpus.

11 http://www.ifri.org/ page consultée le 1er octobre 2017.

4. [Scientext]

Scientext (pages 43 et 71) est un riche assemblage de textes scientifiques et académiques, constitué et mis à disposition par le LIDILEM à travers un site web[12].

Il permet d'interroger notamment des articles scientifiques de diverses sciences humaines et sociales : linguistique, psychologie, sciences de l'éducation, TAL. Il comporte aussi un ensemble de textes scientifiques en anglais, en médecine et en biologie.

Le nombre de textes d'un corpus dépend de la sélection opérée par chaque utilisateur du site.

J'ai utilisé Scientext pour plusieurs recherches : sur les intertitres (section 4.3), sur la formulation en anglais de l'objet d'une recherche (section 5.3).

Une partie de Scientext est en intersection avec le corpus Termith.

5. [Termith]

Le corpus Termith (page 71) reprend certains textes collectés par le projet Scientext et y ajoute des articles scientifiques aspirés sur les sites web de revues de Sciences Humaines et Sociales, afin de couvrir dix disciplines différentes de SHS : linguistique, psychologie, sciences de l'éducation, traitement automatique des langues, anthropologie, géographie, histoire, sciences de l'information et de la communication, sciences politiques, sociologie.

Dans chaque discipline, ce sont 50 articles de recherche qui sont récoltés. Le corpus permet l'étude de la phraséaologie et du lexique scientifique. J'ai contribué aux travaux menés sur le Lexique Scientifique Transdisciplinaire (section 3.2).

6. [EIIDA]

Le corpus EIIDA (page 106), collecté dans le cadre du projet EIIDA, Étude Interdisciplinaire et Interlinguistique du Discours Académique, qui a impliqué des chercheurs de différents laboratoires en France et en Espagne, reste dans l'univers de la production scientifique en rassemblant des textes d'articles scientifiques et des enregistrements de présentations de conférences dans deux domaines, linguistique et géochimie, et dans trois langues, anglais, espagnol, français.

Là encore, c'est le nombre de textes qui a guidé la constitution du corpus : 15 textes pour chaque variable (discipline et mode de communication) dans chacune des langues, d'où un ensemble de 60 textes dans chaque langue – à l'exception de l'espagnol, pour lequel il a été impossible d'obtenir suffisamment d'enregistrements oraux.

Le corpus a été utilisé pour des études contrastives, en particulier entre oral et écrit (section 4.4).

12 http://scientext.msh-alpes.fr/ page consultée le 1er octobre 2017.

7. [Littéracie Avancée] et [Sup-chinois]

Il s'agit de deux corpus d'écrits d'étudiants. Le premier (page 141) réunit des textes de divers genres académiques (mémoires, dossiers, rapports de stage, etc.) rédigés en français dans des universités françaises dans le cadre ordinaire de cursus de sciences du langage ou de formation d'enseignants, de la licence à la maitrise. Il est conçu selon la même logique que Scientext, pour permettre la constitution de corpus en fonction des objectifs de recherche. De par la variété des genres, les 339 textes du corpus dans son état en 2017 peuvent aller de quelques pages à une cinquantaine. Tous les textes ont été réunis au format électronique.

Le second (page 145) a été collecté par Rui Yan dans le cadre de sa thèse sur l'analyse des patrons verbaux de l'écrit scientifique en vue de leur enseignement en Français sur Objectifs Universitaires. Il est constitué de mémoires de master, rédigés en français par des étudiants chinois apprenant le français, dans des disciplines de sciences humaines variées : littérature, sociologie, linguistique.

Les deux corpus ont été utilisés pour une comparaison des compétences des étudiants de niveau master sur le plan de l'écriture académique (section 6.1.2). Ont donc été sélectionnés des textes comparables au niveau de la taille et du genre, c'est-à-dire des mémoires de master.

Partie I
Descriptions linguistiques…

Chapitre 2 - Ancrage théorique, données et méthodes

Se définir comme je l'ai fait précédemment comme linguiste de corpus, mobilisant des outils pour la production des données et pour leur analyse, est un début de positionnement qui doit être complété par une discussion des enjeux et des implications de ce positionnement.

J'aborderai dans une première partie de ce chapitre les questions qu'il me semble crucial d'éclaircir :

- le sens que je mets derrière la notion de « linguistique de corpus » – est-ce une méthodologie, ou une théorie, ou autre chose ? ;
- le statut du corpus dans mes recherches – réservoir d'exemples ou substrat exclusif de l'analyse ? ;
- la conception sous-jacente de la langue.

Quoique je ne résume pas la linguistique de corpus à une méthode ou une méthodologie, les questions de méthode me paraissent mériter une attention au moins aussi importante que l'analyse des données. Dans une seconde partie du chapitre, je poursuivrai donc par une réflexion et des propositions sur les méthodes.

2.1 Linguistiques de corpus, vingt ans après

En (1997) paraissait l'ouvrage de Habert, Nazarenko et Salem qui faisait le point sur les linguistiques de corpus. Cet ouvrage, qui n'était pas un manuel mais plutôt un guide, au sens de ce qui aide à s'orienter, a contribué au début des années 2000 à populariser le recours au corpus et à rendre visibles et lisibles les recherches sur corpus. L'influence du courant anglo-saxon, auquel on doit le terme *corpus linguistics* (Léon, 2008), y est présente dès les premières lignes et surtout dans la définition proposée page 11 pour le corpus, qui est une traduction de la définition avancée par J. Sinclair. Cette influence s'est exercée sur moi aussi, notamment en raison du cadre dans lequel j'ai été formée, et sera sans nul doute sensible dans les positions que je défendrai ici et dans les références que je mobilise.

Cet ouvrage de 1997 fait œuvre utile en balisant le terrain – les terrains – des linguistiques de corpus et en dotant les linguistes d'un bagage technico-pratique pour appréhender la recherche en corpus, mais n'aborde pas les questions épistémologiques qui ont surgi dès le début des années 90 puis dans les années 2000 dans la *corpus linguistics*. Notamment n'y est pas traitée une question que j'ai sentie comme essentielle et qui a guidé ma réflexion dans (Jacques, 2005a) [12][13] : pourquoi une linguistique de corpus, pourquoi, lorsqu'on est linguiste, choisir de travailler sur des corpus plutôt que par introspection ? La première phrase de la conclusion de cet article résume assez bien le propos :

> Nous avons essayé de montrer que ce n'est pas tant sur une question de méthode que la linguistique introspective et la linguistique de corpus divergent, que sur leurs préoccupations respectives et leur conception de la langue. La linguistique de corpus prend sens dans la réintroduction de la question de l'usage, elle amène à situer, c'est-à-dire à replacer les phénomènes observés et décrits dans un contexte. Les possibilités du système sont alors saisies à travers les réalisations effectives, les textes. (Jacques, 2005a : 29) [12]

13 J'indique entre crochets le numéro du texte dans le volume 2, qui rassemble mes travaux publiés.

Ainsi que l'ont souligné Cori et David (2008) et ainsi que cet extrait le met en évidence, la question était double : d'une part cerner la spécificité d'une « linguistique de corpus » par opposition à une autre linguistique (introspective), d'autre part affirmer le programme de recherche particulier de cette linguistique de corpus et, avant tout, la conceptualisation, qui me semble originale, de l'objet de cette recherche.

L'objet principal de (Jacques, 2005a) [12] était d'étayer la thèse selon laquelle le choix d'un travail linguistique sur corpus ne se réduit pas à un choix méthodologique quant à la production des données de la recherche mais répond à certains objectifs de recherche et est cohérent avec une certaine conception de la langue et du travail du linguiste. Il s'agissait alors non seulement de contribuer à l'éclairage des champs et des modes de recherche concernés, mais ce faisant d'invalider divers arguments opposés à ou en faveur de l'une ou l'autre démarche. Je vais dans cette section 2.1 reprendre les divers points qui émergent des argumentations, afin de mettre en évidence les divers enjeux et de clarifier mon propre positionnement.

2.1.1 La production des données

Comme le montrent fort justement Cori et David (2008), nombre d'articles et d'ouvrages de linguistes se réclamant de « l'approche sur corpus » (l'expression est de Cori et David et me paraît plus judicieuse que l'englobante mais fallacieuse dénomination « linguistique de corpus ») dans les années 90 et au début des années 2000 sont émaillés d'observations, généralement mélioratives, sur les bienfaits de la linguistique de corpus comparativement à une linguistique qui ne serait pas de corpus. La spécificité la plus visible de l'approche sur corpus tenant au mode de production des données de la recherche, c'est sous cet angle qu'une autre approche, la linguistique introspective, est prise comme terme comparant pour établir le bien-fondé de l'approche sur corpus. Engagé dans une réflexion sur « la production des données en linguistique introspective », Corbin affirme :

> Deux façons pour un linguiste de constituer les données sur lesquelles il travaille : l'introspection, le corpus. Elles découpent le champ des recherches linguistiques en deux domaines, qu'il est commode de baptiser schématiquement « linguistique de bureau » et « linguistique de terrain », et dont aucun ne peut légitimement être présenté comme incarnant à lui seul LA linguistique : éventuellement complémentaires, voire présentant certaines intersections, ces deux linguistiques ne peuvent pas avoir globalement le même objet. (Corbin, 1980 : 121)

Remarquons avec Cori et David (2008) que ni l'une ni l'autre ne recouvre en fait de champ homogène, c'est d'ailleurs ce qui motivait le pluriel du titre de l'ouvrage de Habert *et al.* (1997), de même, l'hétérogénéité des linguistiques introspectives est patente à la lecture de Corbin (1980), mais malgré cela, se dessine à travers les arguments mobilisés une série de dichotomies qui mettent en opposition deux figures tranchées de linguistes.

Sur la production de données, sont opposés d'un côté les corpus, fournisseurs de données naturelles, authentiques, j'aurais presque envie d'ajouter « non trafiquées », de l'autre l'invention par le linguiste d'exemples sur lesquels raisonner et bâtir la description. Surgit là une première opposition qu'il faut à mon sens s'empresser de déplacer pour la dépasser : les linguistes de corpus défendent l'idée que les corpus fournissent un échantillon fiable de la langue et de son usage, les linguistes de l'introspection leur rétorquent que l'échantillon est biaisé par le corpus qui a été réuni, lequel ne peut donner accès à toute la langue mais seulement à ce que contient le corpus, qui n'est autre chose que ce que le linguiste a décidé d'y mettre. C'est ainsi que Hoek le récusait : « la délimitation d'un corpus décide d'avance des résultats : on n'y trouvera pas

autre chose que ce qu'on y a mis. Une grammaire ne saurait être déduite à partir d'un corpus » (Hoek, 1981 : 19).

De leur côté, les linguistes de corpus (ou de terrain) dénoncent le caractère subjectif, non fiable, non reproductible de l'intuition. Dans un article qui est devenu une référence, Fillmore rapporte une anecdote sur une intuition démentie par les faits :

> A few years ago, my (I think) friend William Labov went around the world giving a lecture in which something that I had written was offered as a paradigm example of what he called "woolly minded introspectionism". In attempting to demonstrate certain kinds of fit between linguistic form and aspects of language use, I had suggested that a particular utterance form could not be used over the telephone. My example involved the colloquial gesture- requiring demonstrative yea, as in *It was about yea big*. For this sentence, the addressee has to be watching the speaker (Fillmore 1972). Labov, master observer of language as he is, soon after reading my claim, heard somebody use just that expression over the telephone. I am convinced that the person Labov heard would have corrected himself instantly if he had realized what he had just said, but nevertheless I stand accused and convicted as a woolly minded introspectionist. (Fillmore, 1992 : 38)

Le sport favori des linguistes de terrain semble ainsi être de traquer les contradictions apportées par la vraie vie aux jugements des linguistes d'introspection. Labov, puisqu'il était question de lui, a systématisé une réflexion sur les limites de l'intuition et arrive à la conclusion qu'elle ne permet pas d'atteindre des données fiables et reproductibles (1975, 1996). Les raisonnements menés sur les données obtenues par introspection sont donc invalidés par cette absence de fiabilité et de reproductibilité, les linguistes procédant par introspection n'atteignant au bout du compte pas le système dont ils prétendent rendre compte mais seulement leur perception de ce système, perception inévitablement informée par leur appartenance à certains groupes régionaux – donc exposés à certaines variantes dialectales et pas à d'autres – et surtout à certains groupes culturels et sociaux, en l'occurrence non seulement exposés aux usages normatifs de la langue mais les ayant probablement intériorisés.

Le bilan que l'on peut tirer de ces critiques de part et d'autre est que si on les suit, la production de données en linguistique est quasiment à coup sûr vouée à la défaillance. À l'incomplétude inévitable des corpus répond la faillibilité de l'introspection, à laquelle s'ajoute le filtre inconscient du linguiste. Laissons à nouveau la parole à Fillmore, qui explique que même le plus petit corpus lui a permis d'entrevoir des faits qu'il n'aurait jamais pu atteindre autrement :

> This paper is a report of an armchair linguist who refuses to give up his old ways but who finds profit in being a consumer of some of the resources that corpus linguists have created.
>
> I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body. (Fillmore, 1992 : 35)

Sur cette question de filtre et de sélection, Cori et David (2008), de même que moi-même (Jacques, 2005a) [12], font remarquer que, malgré l'appellation de *données*, les faits que le

linguiste de corpus analyse ne jaillissent pas spontanément du corpus et ne sont pas traités indistinctement quels qu'ils soient. Les « données » sont tout aussi bien jugées et évaluées, voire filtrées, avant d'être analysées.

Un excellent exemple en est donné par l'étude de Plénat *et al.* (2002). Désireux de recueillir des attestations de mots présentant un suffixe *este*, les auteurs ont conçu un algorithme de recherche pour parcourir le web et enregistrer les mots correspondant au critère déterminé. Comme on peut s'y attendre, à côté de formes jugées satisfaisantes au regard du critère, d'autres ont été écartées comme « coquilles » ou « erreurs ». Si pour certaines, ce statut d'erreur est plutôt incontestable (par exemple *librairieste Thérèse* pour *librairie Ste-Thérèse*), pour d'autres, il y a eu jugement disqualifiant la forme (par exemple, *écologieste* pour *écologiste*). Il ne s'agit pas ici de remettre en cause la décision assurément justifiée de collègues plus avertis que moi sur les questions de morphologie dérivationnelle mais de montrer par cet exemple que le filtre n'opère pas uniquement dans le cas de l'introspection et que tous les linguistes recourent à leur connaissance de la langue et de son fonctionnement à un moment ou un autre de la production des données de la recherche. Et tous, lorsqu'ils écartent certaines formes ou constructions au motif qu'elles relèvent d'erreur, de coquille ou de « raté », le font en toute bonne foi – du moins veux-je le croire – sans avoir aucunement l'impression qu'ils pourraient fausser abusivement le recueil de données.

Ayant compris ceci, chaque linguiste doit mettre en place des méthodes et des procédures pour limiter l'arbitraire et garantir, sinon accroitre la reproductibilité et la vérification de la production des données. Je ferai diverses propositions en ce sens dans la seconde partie de ce chapitre (page 41).

Un second point de divergence entre approche sur corpus et approche par introspection découle du mode de production des données, il recouvre la distinction classique entre science empirique et science expérimentale.

2.1.2 Démarche empirique *vs* démarche expérimentale

L'approche sur corpus est une approche intrinsèquement empirique, elle procède de la même démarche que le botaniste qui collecte des plantes ou que le météorologue qui enregistre jour après jour des données sur le temps. Dans l'introduction de son ouvrage, Tognini-Bonelli résume très bien cette démarche et ses implications :

> Corpus work can be seen as an empirical approach in that, like all types of scientific enquiry, the starting point is actual authentic data. The procedure to describe the data that makes use of a corpus is therefore inductive in that it is statements of a theoretical nature about the language or the culture which are arrived at from observations of the actual instances. The observation of language facts leads to the formulation of a hypothesis to account for these facts; this in turn leads to a generalisation based on the evidence of the repeated patterns in the concordance; the last step is the unification of these observations in a theoretical statement. (Tognini-Bonelli, 2001 : 2)

Un terme important dans cette citation, sur lequel je reviendrai pour me positionner, est le terme « inductif » : une approche empirique des faits linguistiques n'exclut aucunement une théorisation – en ce sens il ne s'agit pas d'opposer empirie à théorie –, mais cette théorisation est **induite** des faits et observations collectés et doit pouvoir rendre compte de faits non observés, dans la mesure où ils sont de même nature (autrement dit de même type) que les faits observés.

La démarche prônée par Tognini-Bonelli (2001) est exclusive dans les deux directions : le linguiste doit rendre compte de tous les faits observés dans le corpus, c'est-à-dire qu'il ne doit pas y avoir de reste dans l'élaboration théorique, et le corpus est le seul fournisseur des faits à prendre en compte, c'est-à-dire qu'il ne doit pas y avoir de connaissances (réelles ou supposées) extérieures au corpus injectées dans la théorisation. Ce dernier point répond à la critique, précédemment évoquée, qui est souvent adressée à l'approche par introspection, de prendre pour LA langue le seul idiolecte du linguiste et de fonder une élaboration théorique sur une représentation individuelle et subjective des faits de langue (très précisément le genre de mésaventure rapportée par Fillmore, voir plus haut).

Cette position fermement empirique se laisse résumer par la formule : *les données, rien que les données, et toutes les données*.

À l'inverse, l'approche par introspection peut être rapprochée d'une démarche expérimentale : il s'agit de prendre un énoncé ou une série d'énoncés et de les manipuler de façon contrôlée et systématique pour observer les effets des manipulations. Corbin rapporte ainsi les propos de Gross :

> Tester l'acceptabilité d'une séquence, c'est procéder à une expérience. La construction d'exemples et de contre-exemples constitue l'activité expérimentale du linguiste qui vérifie la théorie de certains phénomènes (Gross 1975 : 19). […]
>
> « créer des conditions expérimentales favorables » (Gross 1975 : 20) consiste à protéger les raisonnements grammaticaux, lors de la constitution puis de l'exposition d'une argumentation linguistique, contre toute critique portant sur les exemples sur lesquels ils se fondent, en s'assurant du caractère « tranché » (…) de l'*acceptabilité* et de l'absence d'ambigüité sémantique (…) de ces exemples, la possibilité d'une « reproduction des expériences » (…) étant la condition d'existence de toute science expérimentale. (Corbin, 1980 : 136)

Le travail du linguiste qui recourt à l'introspection est ainsi comparable au travail du physicien qui monte une expérience : il élabore un système explicatif pour un ensemble de faits, met en place une situation dans laquelle les prémisses sont contrôlées, introduit le facteur qui doit modifier la situation et vérifie si les changements obtenus réalisent les attentes telles que la théorie formulée les anticipait. L'ensemble de son expérimentation doit être contrôlée et reproductible avec, en principe, les mêmes résultats.

Cori et David reprennent aussi cette idée d'expérimentation, tout en insistant avec Auroux sur la singularité de la linguistique qui n'a pas besoin de l'appareillage des autres sciences expérimentales pour procéder :

> L'exemple présente certaines des propriétés d'une expérimentation. […] Auroux rappelle que l'expérimentation, en la matière, n'est pas chose nouvelle, puisque des protocoles tels que « couper des phrases, permuter ses éléments, etc. » sont connus depuis longtemps. Ce sont des manipulations sans instrument : « L'une des particularités des sciences du langage, en effet, c'est que le langage est sans médiation à disposition du locuteur : je puis produire, à volonté, des phrases, les tronquer, y introduire tel élément que je choisis, etc. Il se pourrait que ce soit le seul exemple d'une manipulation sans instrument, du moins le seul qui se soit maintenu dans un état développé d'une discipline scientifique. » [Auroux 1998 : 170] (Cori & David, 2008)

Si l'on envisage, comme je le propose à la suite de ces auteurs, l'opposition entre approche sur corpus et approche introspective comme une distinction entre science empirique et science

expérimentale, la linguistique se trouve du point de vue scientifique dans la même position que la psychologie : d'un côté l'expérience contrôlée, qui réduit pour les besoins de l'étude les variables et les simplifie au risque d'appauvrir la situation étudiée et d'anéantir ce faisant toute possibilité d'extrapolation aux situations réelles (la vraie vie dans laquelle évoluent les vrais sujets de l'expérience) ; de l'autre le foisonnement des observations authentiques, dans lesquelles tant de paramètres interagissent pour produire les effets observés qu'il est parfois hasardeux d'avancer un système explicatif qui prenne en compte la totalité des facteurs impliqués et qu'il est quasiment impossible de bâtir un système prédictif.

L'analogie avec le climat et la météorologie me parait ici propre à éclairer l'alternative : on connait bien les mécanismes qui produisent la pluie, la neige ou les tornades – le pouvoir descriptif de la météorologie est important – mais pour autant on n'est pas en mesure de prévoir sans la moindre marge d'erreur la survenue de certains phénomènes, il n'est qu'à voir pour s'en convaincre les approximations de certaines alertes météo qui soit ont minimisé les ampleurs des phénomènes survenus, soit au contraire les ont exagérées.

Il me semble que le choix de l'approche sur corpus est de privilégier l'adéquation aux données et de donner ainsi la priorité à la qualité de la description, en d'autres termes, rendre compte de ce que la langue est, sans nécessairement proposer un système prédictif pour tout fait de langue. De même que la météorologie est capable d'indiquer qu'à partir des données observées, plusieurs évolutions du temps sont possibles et parmi celles-là, une ou deux sont plus probables, la linguistique peut faire émerger des possibles et des probables mais ne peut raisonnablement pas prétendre à être aussi fermement prédictive que certaines lois de la physique (encore que la physique ait connu au 20^{e} siècle avec la physique quantique une remise en cause de ses fondamentaux).

Corbin, qui cherche à faire de l'introspection un outil satisfaisant, défend une position plus orientée vers la prédiction :

> *L'introspection peut alors être conçue comme l'instrument privilégié d'une recherche sur les limites ultimes du possible prédictible à partir des observables.* (Corbin, 1980 : 155) [en italique dans le texte original]

À ce point, la question sous-jacente est celle à laquelle j'avais tenté en 2005 d'apporter des éléments de réponse : ce qui distingue les deux approches n'est-il pas plutôt une conception sous-jacente de l'objet d'étude et des objectifs du linguiste qu'un mode de production des données et une démarche d'analyse ?

2.1.3 Objet et objectifs

Est-ce que l'approche sur corpus et l'approche par introspection travaillent sur le même objet ? Une partie de la réponse apportée par les linguistes de corpus est donnée dans un manuel de linguistique de corpus :

> Corpus linguistics today is often understood as being a relatively new approach in linguistics that has to do with the empirical study of “real life” language use with the help of computers and electronic corpora. (Lüdeling & Kytö, 2008 : V)

Il s'agit donc d'étudier l'utilisation de la langue dans la vraie vie, pas de construire un modèle hypothétique de la compétence supposée d'un locuteur-auditeur idéal. Mais cette dernière visée ne convainc pas non plus tous les linguistes recourant à l'introspection, Corbin poursuit ainsi les objectifs qu'il assigne à l'introspection :

> La démarche est donc analogique : on part de données indiscutables et on cherche jusqu'où il est envisageable d'étendre les régularités qu'elles suggèrent. Dans cette perspective, l'objet qu'on se propose d'approcher est certes encore *LA langue*, mais conçue cette fois comme la somme des énoncés dont on peut prédire qu'ils sont productibles par tout ou partie d'une communauté linguistique, et non plus de façon réductrice comme le plus petit dénominateur linguistique supposé commun à cette communauté : la compétence du « locuteur-auditeur idéal », dont Chomsky (1965 : 12) fait « l'objet premier de la théorie linguistique », est interprétée comme *somme* des compétences des locuteurs-auditeurs réels de la communauté, et non plus comme *moyenne* hypothétique de ces compétences, lecture qui n'est que trop répandue. (Corbin, 1980 : 155)

Je soulignerai avec un peu de malice qu'en évoquant des locuteurs-auditeurs ***réels*** et en se démarquant de l'idéal, Corbin fait le pas vers l'approche sur corpus qui forme trente ans plus tard une part de ses recherches[14].

L'articulation entre les productions, c'est-à-dire l'usage de la langue, qui forment les observables à analyser, et ce qui derrière (ou au-dessus, selon la métaphore spatiale que l'on préfère) ces productions fait système, qui donc serait LA langue, semble être le point problématique de l'approche sur corpus. Sauf à déplacer la question en récusant cette dichotomie très saussurienne, ainsi que le propose Halliday, à travers une analogie avec le climat :

> We are so accustomed to thinking about language and text in terms of dichotomies such as the Saussurean langue and parole, or Hjelmslev's system and process, that we tend to objectify the distinction: there is language as a system, an abstract potential, and there are spoken and written texts, which are instances of language in use. But the "system" and the "instance" are not two distinct phenomena. There is only one phenomenon here, the phenomenon of language: what we have are two different observers, looking at this phenomenon from different depths in time. If I may use once again the analogy drawn from the weather: the instance-observer is the weatherman, whose texts are the day-to-day weather patterns displaying variations in temperature, humidity, air pressure, wind direction and so on, all of which can be observed, recorded and measured. The system-observer is the climatologist, who models the total potential of a given climatic zone in terms of overall probabilities. What appears to the former as a long-term weather pattern becomes for the latter a defined climatic system. There is only one set of phenomena here: the meteorological processes of precipitation, movement of air masses and the like, which we observe in close-up, as text, or else in depth, as system. But one thing is clear: the more weather we observe, as instance-watchers, the better we shall perform as system-watchers when we turn to explaining the climate. (Halliday, 1992 : 66)

Le « système » ne serait pas autre chose qu'un changement d'échelle dans l'appréhension des phénomènes et il serait tout à fait valide de penser l'atteindre en collectant et assemblant les observations individuelles (les textes).

Stubbs (2001 : 243), à travers une autre analogie, va à peu près dans la même direction, en indiquant que le problème posé à la linguistique est au bout du compte un problème récurrent dans les sciences empiriques : la relation supposée entre des produits (observables) et un

14 Il a en effet constitué des ressources pour l'étude de la langue du football, en collaboration avec N. Gasiglia (2004).

processus (non directement observable). Il compare la linguistique à la géologie parce que cette dernière discipline est confrontée à la difficulté de retracer, à partir d'états de roches et de formations géologiques, un processus non atteignable du fait qu'il s'étend sur une période temporelle excédant l'expérience humaine directe. C'est cependant ce processus qui intéresse la géologie. D'une façon similaire, la linguistique est intéressée par un processus lui aussi inobservable, mais dans ce cas parce qu'il « s'étend » sur de nombreux locuteurs différents. Et de même que les produits géologiques peuvent avoir été influencés par l'environnement (climat, par exemple) tout en restant tout de même gouvernés par les mêmes grands processus généraux (érosion et sédimentation, par exemple), les produits linguistiques peuvent être influencés par l'environnement (le contexte sociolinguistique) mais restent gouvernés par de grands processus généraux (pour Stubbs, collocation et colligation, cette dernière étant une collocation au niveau grammatical et non simplement au niveau lexical).

Il ressort de cet examen un mouvement finalement similaire pour les deux grands types d'approches, qui me conduit à tenter de compléter et de mieux formuler ce que j'ai écrit en 2005. L'objectif de l'approche sur corpus n'est pas limité aux usages, il concerne bien l'appréhension de la langue comme système. C'est plutôt du côté de ce qu'est la langue, de ce qui la gouverne, et de sa place relativement aux autres composantes de ce qui fait de l'humain un humain que passe la ligne de partage entre une linguistique qui se fonde sur les corpus et une linguistique qui bâtit et manipule des exemples. Je ferai ici référence à Chafe (1992) qui oppose deux visions de la langue, desquelles découlent deux façons de concevoir et de mettre en œuvre le travail du linguiste. La langue peut être considérée au sein des aptitudes humaines comme un système autonome, « un module langagier indépendant » (on reconnaitra là une allusion à Chomsky et à ses propositions), ce qui implique :

> Linguists with this belief feel joy whenever they discover a linguistic phenomenon they can characterize as arbitrary and unmotivated - one they can assign to the independent language module. (Chafe, 1992 : 80)

Mais il peut aussi être considéré que la langue est une part inséparable de l'activité mentale et prend place au sein de l'ensemble des fonctions cognitives, avec lesquelles elle est en étroite dépendance – on aura reconnu là la position des fonctionnalistes tels que M.A.K. Halliday. Dans ce paradigme,

> [l]inguists who adhere to this integrated view feel joy whenever they discover a way in which some linguistic phenomenon can be characterized as motivated and functional - explainable within a larger, coherent picture of the mind. (Chafe, 1992 : 81)

La dualité ici évoquée projette des corrélations en termes de méthodes. Pour mettre en évidence les propriétés d'un système indépendamment de son utilisation et des fonctions qu'il remplit, expérimentations et manipulations sont tout à fait appropriées. On peut par exemple étudier la mécanique et les propriétés des corps en mouvement sans nécessairement envisager une machine particulière.

C'est d'ailleurs dans cette logique de dissociation du système et de son usage que se serait inscrit Chomsky, qui, d'après Léon, ne disqualifiait pas les études de corpus et le développement de modèles probabilistes de l'usage dès lors que la syntaxe n'était pas la cible :

> Given the grammar of a language, one can study the use of the langage statistically in various ways; and the development of probabilistic models for the use of language

> (as distinct from the syntactic structure of language) can be quite rewarding. (Chomsky, 1957, note 4: 17). Cité par (Léon, 2005 : 45)

Mais le problème qui se pose au linguiste – comme à nombre de scientifiques des sciences humaines – est double : d'une part, la déconnexion d'avec les situations authentiques conduit à une artificialité des résultats (l'inadéquation pragmatique de certains énoncés jette un doute sur la validité des théories élaborées à partir de leur examen), d'autre part, les résultats obtenus sont finalement assez pauvres et peu informatifs, c'est la critique assez radicale formulée par Buyssens (1969) à la réception des propositions de Chomsky, médiées en France à la fin des années 60 par N. Ruwet. Pour Buyssens, soit Chomsky est dans l'erreur, soit ses conclusions n'apportent pas grand-chose de plus que la tradition grammaticale qui l'a précédé. Il conclut par ces mots : « La linguistique n'a que faire de la discipline déductive de Chomsky ; elle souffre d'un recours insuffisant à la méthode inductive. » (Buyssens, 1969 : 857).

En effet, dès lors que le langage est pensé dans sa dimension sociale, cognitive et psychologique, la description linguistique doit obéir à un principe impérieux d'adéquation aux données et de plausibilité – un peu comme lorsque, ayant établi les lois du mouvement, on envisage le fait qu'il prend place dans un fluide (air ou eau, par exemple), lequel introduira des contraintes susceptibles d'amender les lois précédemment établies. L'adéquation aux données ne sera garantie que par le rassemblement d'un nombre suffisant de données et par l'établissement de régularités, l'approche sur corpus est ainsi cohérente avec cette conception.

2.1.4 Linguistique de corpus : discipline à part entière ?

Il ressort de ce qui précède que l'on ne peut réduire l'approche sur corpus à une méthodologie – quoique la question soit controversée parmi les linguistes de corpus eux-mêmes. Taylor (2008) examine précisément la façon dont les auteurs d'articles publiés dans divers journaux et affichant les mots-clés *Corpus linguistics* se positionnent.

Pour Aarts – à qui l'on attribue la formation de la dénomination *Corpus linguistics* –, comme pour Teubert (2005) ou Williams (2006), la linguistique de corpus est une discipline. Leech (1992) va plus loin en évoquant une nouvelle approche philosophique et un nouveau paradigme. De même, Stubbs (2001) rejette la définition de la linguistique de corpus comme une simple méthodologie et Teubert (2005) la considère une approche théorique de l'étude de la langue.

Ces positions sont assez cohérentes avec ce que j'ai développé dans la section précédente : si en effet, l'approche sur corpus accompagne une certaine conception de la langue, des objets et des objectifs de la linguistique, alors il n'est pas exagéré de lui proclamer une portée théorique et d'en faire un paradigme de recherche.

Quelques voix ne s'expriment toutefois pas à l'unisson de cette position. Pour McEnery *et al.* (2006), la linguistique de corpus est un système global de principes et de méthodes pour l'utilisation des corpus dans l'étude de la langue, nulle prétention à la constituer en théorie ou en discipline. En 2012, Gries maintient cette position et écrit :

> Put differently, I don't accord CL [corpus linguistics] the status of a theory just as I don't think there is a linguistic theory called experimental linguistics or self-paced reading time linguistics even though, just like results from CL, results from self-paced reading times may call into question units/structures/processes assumed in the kind of formal linguistics that (some of) CL was a reaction against. (Gries, 2012 : 43)

Gries suggère un lien entre la revendication à être une discipline, avec un ancrage théorique, ou à être une méthodologie et la façon d'exploiter le/les corpus. Plus l'approche serait guidée par le corpus (en anglais *corpus-driven*), sans préconception ou hypothèse préalable de ce qu'il va livrer et du fonctionnement des unités que l'on va y étudier, plus marqué serait le positionnement comme théorie. À l'inverse, plus l'approche serait du type « basée sur le corpus » (en anglais *corpus-based*), plus marqué serait le positionnement comme méthodologie.

Il est intéressant de remarquer que Tognini-Bonelli (2001), qui a établi cette opposition entre *corpus-driven* et *corpus-based* et qui défend fermement une approche *corpus-driven*, prend en compte les réticences avancées quant au statut théorique de la linguistique de corpus et articule précisément renouveau méthodologique (du moins ce qui est présenté par tous les laudateurs enthousiastes de la linguistique de corpus comme un renouveau) et prétentions à un statut théorique. Dès la première page de son ouvrage – ce qui montre à quel point la question restait brulante –, elle place la linguistique de corpus sous le chapeau de la linguistique appliquée, sans toutefois argumenter ce rattachement. Cependant, contrairement à d'autres disciplines qui peuvent se ranger sous le même chapeau (Tognini-Bonelli ne précise pas lesquelles), la linguistique de corpus serait une *pre-application methodology*, en ce que, contrairement à d'autres applications qui partent de certains faits établis et les traite avec les ensembles de règles qui constituent une méthodologie, elle est en position d'élaborer ses propres ensembles de règles et ses connaissances avant même l'application, ce qui conduit le linguiste à élaborer de nouveaux paramètres pour traiter les données et ce qui entraine des changements au niveau des unités considérées. Tognini-Bonelli conclut en faveur d'un statut théorique de la linguistique de corpus.

Je dois avouer quelques difficultés à suivre ce raisonnement car il me parait procéder par deux coups de force successifs : le premier est de ranger sans autre forme de procès ou de justification la linguistique de corpus sous l'étiquette « linguistique appliquée » et d'en tirer une quelconque conclusion, le second est de considérer, là encore sans discussion ni explicitation, qu'une méthodologie est l'utilisation d'un ensemble de règles ou de connaissances dans une situation donnée – afin que l'on puisse juger sur pièce, voici l'extrait où tout ceci est explicité :

> In this context we take the view that although corpus linguistics belongs to the sphere of applied linguistics, it differs from other partner disciplines under the same umbrella in that it can be seen as a *pre-application methodology*. While a methodology can be defined as the use of a given set of rules or pieces of knowledge in a certain situation, by "pre-application" we mean that, unlike other applications that start by accepting certain facts as *given*, corpus linguistics is in a position to define its own sets of rules and pieces of knowledge *before* they are applied; this leads the linguist to make use of some new parameters to account for the data, and this entails a change in what can be referred as the *unit of currency* for linguistic description, corpus linguistics has, therefore, a theoretical status and because of this it is in a position to contribute specifically to other applications. (Tognini-Bonelli, 2001 : 1)

On peut effectivement dire que l'approche sur corpus, de même que tout recueil et tout examen de données « écologiques », authentiques, en confrontant le chercheur à la complexité des faits réels, non simplifiés comme ils le sont dans les expérimentations, peut le conduire, voire même le contraindre à d'autres problématisations, à une redéfinition des unités à prendre en considération et de leurs interactions, ce qui peut déboucher sur de nouvelles théorisations. Mais ce n'est pas le corpus qui mécaniquement fait cela, il est aussi possible de travailler sur corpus dans l'idée de vérifier une théorie élaborée au préalable. En somme, l'opposition n'est

pas réductible à « corpus *vs* introspection », elle est plus complexe que cela et recouvre plutôt l'opposition hypothético-déductif / inductif, sans se dissoudre totalement non plus dans ces deux termes.

On peut d'ailleurs comprendre de cette façon l'assimilation que fait Gries entre *corpus-driven* et revendication d'un statut théorique : le fait de se laisser guider par les faits dans une démarche inductive déplace le moment de théorisation et tend à subordonner la possibilité même d'une théorisation au mode de recueil de données.

Dans ma propre pratique de recherche, j'ai en quelque sorte mis en actes cette position, comme on le verra dans les chapitres suivants. Mais le souci de se laisser guider par les données me paraît requérir une certaine prudence sur deux plans : la supposée neutralité du chercheur, qui devrait aborder les données sans préconception aucune – et je voudrais dans la seconde partie du chapitre argumenter en faveur de l'idée que la constitution même d'un corpus contredit totalement l'absence de préconception – ; la façon de recueillir et traiter les données – et je voudrais là développer un panel de propositions pour une approche sur corpus plus « scientifique ».

À titre de synthèse-bilan pour cette première partie de chapitre, la recherche linguistique que je mène :

- prend sens au regard d'une conception de la langue comme partie intégrante du fonctionnement de l'esprit humain, dans un acte de communication et/ou d'élaboration de connaissances – on verra au fil de ce document que je me réfère volontiers au travail et aux propositions de M.A.K. Halliday ;

- vise à dégager des régularités, en s'inscrivant dans une démarche inductive ;

- doit répondre à une exigence d'adéquation aux données.

La cohérence avec moi-même indique une démarche *corpus-driven*, que j'ai effectivement mise en œuvre dans mes recherches, comme on le verra dans les chapitres suivants. Cette démarche ne se prétend toutefois pas dégagée du savoir et des représentations sur la langue que le simple fait de connaitre sa langue construit. Mais, comme l'indique aussi Tognini-Bonelli (2001 : 91), il ne s'agit pas d'évacuer sa propre perception et sa connaissance de la langue, mais de ne pas en rester prisonnier et d'être capable d'opérer toutes les révisions que la confrontation aux données implique. Un enjeu ici est d'objectiver son matériau d'étude. Le/la linguiste doit faire des choix éclairés et explicites, doit savoir ce qu'il/elle fait et ce qu'il/elle regarde quand il travaille sur corpus et qu'il veut procéder par induction.

2.2 Questions de méthode

Je voudrais ici dans un premier temps revenir sur la question posée par l'approche *corpus-driven* quant à la place de la théorisation. Une démarche inductive suppose un schéma orienté des données vers la construction d'un système explicatif de ces données. De ce fait, le rassemblement du matériau d'étude requiert une attention particulière, sur laquelle je reviens maintenant, en reprenant diverses réflexions et propositions élaborées dans (Jacques & Poibeau, 2010) [21].

2.2.1 Encore la production de données…

Il ne s'agit plus ici de discuter d'un mode de production de données dans le cadre d'une distinction corpus / introspection mais bien, dans la perspective d'une approche sur corpus, de

problématiser la production de données issues d'un corpus en attirant l'attention sur ce qui me parait être de fausses évidences.

Une première question à traiter est celle du corpus lui-même et de sa constitution. Cette question a considérablement évolué en 20 ans, grâce en partie à l'irruption de l'informatique dans quantité de sphères, tant professionnelles que de loisirs, et donc à l'existence et à la disponibilité potentielle de nombreux écrits sous un format électronique. Même si certains types de textes ne sont toujours pas aisément « rassemblables » en corpus parce que produits quasi exclusivement hors support informatique – je pense notamment aux textes d'apprenants, que j'évoquerai dans le dernier chapitre de ce document d'habilitation – la constitution d'un corpus pose de moins en moins de difficultés techniques.

Il me faut ici rappeler un certain nombre de précisions quasi-terminologiques, qui devront beaucoup à un chapitre d'ouvrage rédigé en commun avec F. Grossmann et A. Tutin et portant sur les corpus écrits dans la linguistique française – exercice agréable et stimulant que cette mise en commun de nos conceptions et connaissances sur le sujet[15]. Il convient de distinguer plusieurs sortes de corpus, même en se restreignant aux corpus monolingues. Ce que l'on appelle les corpus de référence, dont les prototypes sont les corpus élaborés dans le monde anglo-saxon : le *Survey of English Language* de R. Quirk ou le *Brown Corpus* de Kucera et Francis (Léon, 2005), sont des corpus rassemblant une variété la plus diversifiée possible de productions écologiques (c'est-à-dire produites dans des situations authentiques d'utilisation de la langue et non provoquées par le chercheur pour la constitution du corpus) aux fins de fournir une information en profondeur sur une langue donnée. C'est sur de tels corpus étendus et ambitionnant de couvrir « la langue » que se basent des travaux des années 1990 qui ont contribué à mettre la linguistique de corpus sur le devant de la scène : ceux de D. Biber et ses collègues, par exemple, qui élaborent *The Longman grammar of spoken and written English*, ou ceux de J. Sinclair, qui théorise notamment l'exploitation des corpus pour la lexicographie et mène dans ce cadre le projet CoBuild, de construction de ressources dictionnairiques à partir de corpus. Le propos de ces travaux est d'offrir une description de la langue qui soit la plus proche possible de la réalité quotidienne de la langue et qui corrige les défauts maintes fois observés des ressources linguistiques disponibles : peu d'informations sur la combinatoire, exemples tous mis sur le même plan sans prise en compte des phénomènes de fréquence, caractère artificiel des énoncés proposés pour illustrer les points de langue discutés…

Mon propos n'est pas de discuter plus avant de ces corpus, je les mentionne en fait essentiellement pour m'en distinguer. En effet, ainsi que je le développerai dans les chapitres suivants, je ne me suis pas intéressée à « la langue » mais à certains de ses usages bien particuliers, ce qui m'a conduite vers un autre type de corpus, qu'A. Tutin, dans le chapitre commun que j'évoquais plus haut, a nommés *corpus spécialisés*. Ce sont des corpus élaborés par les chercheurs pour une question de recherche précise. Les études linguistiques ont vu depuis le début des années 2000 en éclore quantité. Par exemple, pour sa thèse sur la définition « naturelle », J. Rebeyrolle (2000) a compilé un corpus de textes encyclopédiques et de manuels ; pour son analyse de la métonymie, M. Lecolle (2003) a rassemblé un corpus de textes journalistiques ; pour sa thèse sur l'antonomase du nom propre, S. Leroy (2001) travaille elle aussi sur des articles de presse.

Ces quelques exemples ont en commun d'illustrer parfaitement la façon dont le linguiste de corpus mobilise déjà sa connaissance, sa représentation, son intuition de la langue avant même

15 Il est en cours de rédaction au moment où j'écris ces lignes mais nous l'espèrons en cours de publication en 2018.

d'avoir commencé à examiner des données. Rassembler un corpus en vue de l'étude d'un certain phénomène, c'est déjà faire l'hypothèse que les textes ou le matériau choisis présenteront des occurrences du phénomène en question, c'est donc déjà une représentation du fonctionnement de la langue. Cela peut paraitre une évidence, mais lorsque l'on décide d'utiliser des encyclopédies et des manuels pour « s'intéresser aux définitions telles qu'elles sont spontanément formulées par les locuteurs eux-mêmes pour expliciter le sens des mots qu'ils emploient » (Rebeyrolle, 2000 : 1), on suppose – avec raison ! – que manuels et textes encyclopédiques recèleront des définitions spontanées. De la même manière, le locuteur d'une langue sait ou croit savoir que les textes journalistiques font usage de certaines figures de rhétoriques, que les textes académiques ou scientifiques font usage de citations, etc. Dans cette perspective, l'idée d'une approche exclusivement *corpus-driven* dans laquelle le linguiste suspendrait sa connaissance de la langue pour faire émerger la théorie du corpus lui-même est un leurre et constitue précisément le défaut à partir duquel les contempteurs de l'approche sur corpus ont argumenté leur réfutation de cette approche (voir 2.1.1). Il faut au contraire être pleinement conscient des choix opérés et les justifier par des critères explicites, préférentiellement homogènes (Péry-Woodley, 2001).

Les exemples que j'ai donnés précédemment mettent en lumière une tendance actuelle de la linguistique de corpus : la prise en compte des genres textuels dans leur interaction avec le fonctionnement de la langue. Divers travaux, par exemple (Adam, 2004 ; Branca-Rosoff, 1999 ; Malrieu, 2004), s'attachent précisément à mettre en évidence l'importance du genre à l'égard du fonctionnement du discours et par là même une surdétermination du genre sur le fonctionnement linguistique. Il resterait toutefois, dans la lignée de (Malrieu & Rastier, 2001), à évaluer de façon fine cette interrelation entre genres et caractéristiques linguistiques des textes. C'est une question que j'ai abordée dans (Jacques & Aussenac-Gilles, 2006) [16] à l'occasion de travaux sur le repérage automatique de relations conceptuelles, je reviendrai sur ces travaux dans le chapitre 5 de ce mémoire d'habilitation.

J'espère à ce point avoir convenablement montré la prudence dont il convient de faire preuve au moment de la constitution d'un corpus : non seulement est-il nécessaire de définir très précisément ce pourquoi le corpus est rassemblé, mais encore faut-il peser avec soin chaque élément à y inclure ou à en exclure, expliciter les critères mobilisés et les justifier dans un système homogène.

Cependant, la constitution du corpus n'est que la première étape – et presque la moindre – vers l'établissement du matériau d'étude. L'étape cruciale est celle du recueil des données elles-mêmes. Pour l'illustrer, une petite expérience, clin d'œil à mon collègue (et présentement mentor) F. Grossmann.

Imaginons la question de recherche suivante : l'utilisation rhétorique de la forme *voir* dans l'article scientifique en sciences humaines. Le corpus est presque déjà prêt : Scientext[16] est une base de textes scientifiques et académiques de diverses disciplines, à partir de laquelle il est aisé de sélectionner un corpus selon le genre (article de revue, texte de colloque, thèse ou HDR) et/ou la discipline. Restreignons-nous pour l'expérience aux articles de revues de quatre disciplines rangées dans la catégorie 'sciences humaines' : linguistique, psychologie, sciences de l'éducation, TAL, ce qui constitue un ensemble de 22 textes. Demandons à Scienquest, l'outil de recherche dans Scientext, d'afficher la concordance de la forme *voir*, on obtient 108 occurrences. La première question qui se pose est : toutes ces occurrences sont-elles « bonnes

16 http://scientext.msh-alpes.fr/scientext-site/spip.php?article1

à prendre » ? Si ma question de recherche concerne l'utilisation rhétorique de « voir » dans la construction du discours scientifique dans ces articles, alors le contexte (4) ne « m'intéresse » pas, dans la mesure où il décrit un logiciel mais ne s'insère pas dans le raisonnement scientifique :

(4) Ainsi, dans la continuité de sa lecture du texte, le lecteur se voit proposer, par une signalétique spécifique, des parcours spécifiques sans rupture de la cohésion textuelle puisqu'il peut **voir** à tout instant le texte complet, ce qui lui permet entre autres d'assurer la continuité référentielle[17] [Scientext - TAL]

La question de recherche du chercheur le conduit assez inévitablement à « trier » les données et à écarter celles qui ne lui paraissent pas pertinentes pour son étude.

Est-ce un problème ? Quand la recherche (ici délibérément simplifiée) porte sur des formes et qu'il s'agit de faire un tri, chaque chercheur engagé dans cette démarche se croit de bonne foi bien-fondé à écarter les occurrences qui ne correspondent pas à son critère de sélection et que le logiciel – qui n'accède pas au sens – lui a présentées sur la simple base d'une unification formelle : je demande les occurrences de *voir*, je récupère tous les *voir* d'un texte, même ceux qui correspondent à une erreur orthographique et qui sont en fait *voire*[18], j'élimine ces erreurs et les occurrences qui n'entrent pas dans mon champ de recherche (de la même manière que Plénat *et al.* (2002) écartaient « écologieste »).

Les problèmes potentiels de cette sélection viendraient de deux directions :

1. pour certaines formes, il se peut que la décision d'inclusion ou d'exclusion ne soit pas si nette que cela et qu'elle soit donc variable selon les individus ;

2. quand la recherche n'est pas sémasiologique mais onomasiologique, c'est-à-dire part des significations pour aller vers les formes, la mobilisation de l'interprétation – et donc de la subjectivité du chercheur – est encore plus manifeste et accroit donc la variabilité selon le chercheur.

2.2.2 Variations interindividuelles dans l'appréciation des données

La recherche que je vais évoquer ici (Jacques & Poibeau, 2010) [21] a été rendue possible par mon implication, en tant que chercheure « post-doct », dans le projet ANR Textcoop, au LIPN (Laboratoire d'Informatique de Paris-Nord). Le cadre de ce projet a permis de mobiliser plusieurs chercheurs autour d'un même objet : la caractérisation, en vue d'un repérage automatique, de textes procéduraux. Il a permis en outre de recruter un étudiant en stage de M2 du Master PluriTAL (Université Paris Nanterre). Ces précisions vont permettre de comprendre certains des résultats.

Les principaux acquis de cette recherche concernent les aspects méthodologiques et les réflexions qui les ont accompagnés. Pour leur compréhension, il faut savoir que l'un des objectifs était d'être en mesure de classer automatiquement des textes aspirés sur le web pour distinguer ceux qui parmi eux comportaient des procédures, des consignes. Donc, après l'étape de constitution de deux corpus par aspiration de certains sites web, l'un dans le domaine médical, l'autre dans le champ du bricolage, venait l'étape de recueil des données proprement

17 Couto J., Minel J.-L. (2006) Navigation textuelle : représentation des textes et des connaissances, *Traitement Automatique des Langues*, vol 47, n°2.

18 Certaines études, telles que celles qui seront évoquées ensuite, de J. Rebeyrolle (2000) ou de S. Leroy (2001), ont d'ailleurs pour objectif de raffiner les requêtes afin de permettre aux outils automatiques de cerner les contextes recherchés. La mise au point des requêtes devient une recherche en elle-même, aspect que j'ai traité dans (Jacques, 2016b) [30].

dites, c'est-à-dire des passages ou plutôt segments procéduraux, en vue de leur modélisation. Une première définition, prise comme démarrage de la recherche, des procédures et des segments de textes qui en contiennent s'est appuyée sur l'idée qu'une procédure est une instruction ou une série d'instructions, donc un segment procédural serait un segment du texte à partir duquel on peut effectuer une action[19].

C'est là qu'intervient de façon très cruciale la notion de variation. L'objectif étant de collecter des segments procéduraux et, de façon complémentaire, d'en distinguer les segments non procéduraux, il était nécessaire de :

- prendre comme point de départ une définition des segments procéduraux qui ne s'appuie pas sur leurs caractéristiques linguistiques, puisqu'au bout du compte ce sont ces caractéristiques que la recherche voulait délimiter ; donc,

- prendre comme point de départ un jugement appuyé sur la compréhension (et l'interprétation) des textes – critère potentiellement périlleux en raison de la part de subjectivité qu'il implique ;

- savoir très précisément dans quelle mesure et selon quelle ampleur joue la subjectivité dans la sélection des passages procéduraux.

En effet, dans ce genre de recherche, où la production des données qui seront soumises à l'analyse repose sur un jugement de l'analyste, si ces jugements varient, l'ensemble de données obtenu après sélection sera lui aussi variable et, partant, la description produite de même. On revient finalement aux mêmes écueils d'une description très dépendante de l'intuition du chercheur que précisément le recours au corpus est supposé éviter (comme je l'ai développé en 2.1.1).

La façon de résoudre le problème qui a été choisie a reposé sur la mise à contribution des chercheurs impliqués et du stagiaire M2. Nous avons eu à juger des segments des mêmes textes dans chacun des deux corpus. C'est ensuite une mesure courante dans ce type de recherches en linguistique, le coefficient Kappa (Carletta, 1996), qui a été utilisée pour estimer le taux d'accord entre nos divers jugements.

Les résultats de cette expérience se sont avérés concluants, dans la mesure où ils ont mis en évidence une variabilité interindividuelle selon le corpus. En effet, le corpus du domaine médical, constitué de recommandations issues de l'état de la recherche en médecine, élaborées en vue de guider les médecins vers de bonnes pratiques thérapeutiques, n'est pas clairement injonctif et procédural. La dénomination du genre est assez révélatrice : les documents sont des *recommandations*, pas des recettes, et la recommandation n'adopte pas l'impératif, elle camoufle ses injonctions derrière des formulations telles que « une chimiothérapie permet d'obtenir un taux élevé de rémissions complètes » qui, tout à la fois, donnent au praticien une indication sur le traitement à envisager et le laissent juge de la thérapeutique à mettre en œuvre.

Pour ce corpus, l'accord inter-annotateurs calculé est bon, sans plus, avec un coefficient de 0,72.

En revanche, pour l'autre corpus, constitué de fiches de bricolages, l'accord est bien meilleur, avec un coefficient de 0,85. Ces fiches sont en effet explicitement destinées à permettre au lecteur une réalisation concrète et se rapprochent fortement de la recette de cuisine en énumérant une série d'actions, comme par exemple en (5), extrait d'une fiche pour « Jointoyer un carrelage mural ».

19 Les parangons des textes procéduraux sont par exemple les recettes de cuisine ou les modes d'emploi (Adam, 2001).

(5) Étalez la pâte avec une raclette en caoutchouc en dessinant des "8". Bien garnir les joints. Ils doivent être réguliers et lissés, sans être creusés. [Bricolage]

Toutefois, même pour ces documents où l'expression des procédures suit une forme plus canonique, on voit que l'accord n'était pas parfait, certains segments donnant lieu à des jugements divergents, comme (6), extrait d'une fiche pour l'entretien des bicyclettes.

(6) Dans tous les cas, il est nécessaire de faire réviser la bicyclette par un spécialiste tous les ans mais vous devez être capable de détecter les dysfonctionnements de votre bicyclette et d'effectuer l'entretien minimal. [Bricolage]

Pour l'un des annotateurs, « il est nécessaire de faire réviser » a été jugé équivalent à « faites réviser », donc injonctif, alors que l'autre annotateur n'y a pas vu de guidage pour une action.

En définitive, même pour les cas que l'on pourrait penser les moins sujets à caution, on s'aperçoit que les appréciations des chercheurs divergent, dès lors qu'ils se penchent sur les mêmes données, et même en ayant connaissance de leur contexte – l'explication des divergences dans les jugements d'acceptabilité repose en effet souvent sur l'hypothèse que les chercheurs ne restituent pas les mêmes contextes pour les énoncés à juger, et que c'est en fonction de leur capacité à imaginer un contexte pertinent qu'ils acceptent ou non l'énoncé proposé ; rien de tel ici, les chercheurs avaient accès au texte dans son intégralité.

Cette variabilité peut être considérée non comme un défaut ou un problème, mais au contraire comme une source de connaissances supplémentaires dans le processus de recherche. En effet, si on regarde à nouveau la démarche introspective, dont j'ai précédemment mis en lumière le caractère « expérimental », ses tenants défendent comme une richesse inestimable la possibilité de pouvoir forger des exemples contrefactuels. Il s'agit, pour reprendre les propos de Corbin (1980), d'explorer les limites du système en contrastant le possible et l'impossible, ce que le système permet et ce qu'il ne permet pas. L'approche sur corpus n'est confrontée qu'à ce qui a été réalisé ; elle ne peut affirmer que ce qui n'est pas rencontré est impossible, elle peut juste indiquer que cela n'a pas été rencontré.

Nous[20] proposons de voir dans la confrontation de jugements sur les données un moyen de mettre en évidence ce qui semble central et ce qui semble périphérique par rapport au système : les occurrences sur lesquelles les jugements divergent sont les occurrences qui ne sont pas « typiques » du phénomène étudié mais en manifestent les marges, les limites. Elles sont donc tout aussi riches d'informations pour la compréhension du phénomène que les énoncés fabriqués assortis d'astérisques dans la démarche introspective. Il en découle que, au-delà de la production des données, l'examen par plusieurs chercheurs des mêmes occurrences est déjà un temps de recherche et d'analyse.

La recherche évoquée, et les évolutions récentes de la constitution et de la mise à disposition de corpus en France me conduisent à formuler des propositions concrètes, dont la mise en œuvre n'est peut-être pas réaliste, malheureusement, étant donné les conditions matérielles des recherches linguistiques, mais que je prends comme guide méthodologique pour mes propres recherches.

2.2.3 Des propositions concrètes (mais peut-être guère réalistes)

J'ai récemment repris ces thématiques dans une réflexion (Jacques, 2016b) [30] qui s'est enrichie de ma participation active, depuis 2011, au consortium Corpus Ecrits du TGIR Huma-

20 « Nous » désigne ici Thierry Poibeau et moi-même (2010) [21].

Num, consortium qui en 2016 a fusionné avec un autre consortium d'Huma-Num, « Corpus Oraux et Multimodaux », ce qui a engendré le consortium Corpus, Langues et Interactions (CORLI).

La participation à ce consortium m'a permis de mieux cerner l'état de la recherche sur corpus en France et même si ce consortium ne constitue pas un observatoire systématique des recherches linguistiques se fondant sur des corpus, il en offre un aperçu.

L'un des objectifs du consortium est la publication et la diffusion des corpus constitués isolément par les chercheurs en linguistique. Nombre de projets de recherche impliquent en effet l'élaboration d'un corpus (qui entre dans le champ des *corpus spécialisés*) ou/et la production de ressources, qui n'ont en général guère de diffusion au-delà du laboratoire ou du projet qui les ont vu naitre, sauf démarche volontaire et délibérée du laboratoire porteur – par exemple, la page « REDAC » de l'ERSS[21] est un modèle du genre : elle donne accès à des corpus libre de droits, des outils, des lexiques… Certains corpus qui seraient d'une grande valeur pour l'ensemble de la communauté restent confinés aux « tiroirs » ou aux « disques durs » des doctorants qui les ont élaborés, faute d'un appui suffisant pour leur mise à disposition, qui est en fait exigeante : négociation des droits pour la diffusion, définition du ou des formats de diffusion, définition des métadonnées assurant la visibilité du corpus, réalisation technique, suivi… Constituer et travailler un corpus « pour soi », dans le cadre limité de sa propre recherche, est sans commune mesure en terme d'investissement en temps, en énergie et en connaissances techniques et pratiques avec la constitution d'un corpus publiable et diffusable à l'ensemble de la recherche.

Cette difficulté à partager les données primaires de la recherche est à mon sens un verrou, à divers niveaux. Je laisse ici la parole à T. Chanier, qui a très bien exposé, dans le rapport final du projet ANR Mulce, les limitations dues au caractère confidentiel des corpus mobilisés dans les recherches – il s'agissait pour Mulce de recherches liées aux apprentissages en ligne, mais les remarques qui suivent sont tout à fait transférables aux études linguistiques sur corpus :

> Le bilan qui a été fait […] en janvier 2005 à Amiens [...] montre que les travaux des domaines se rapportant à l'utilisation des technologies dans l'apprentissage (EIAH, TICE, CSCL, AL &SIC) sont assez peu reconnus scientifiquement pour différentes raisons : les chercheurs en psychologie ou en sciences de l'éducation reprochent bien souvent le manque de méthodologie dans le protocole d'expérimentation, d'autres le manque de réplicabilité ou de validité des résultats avancés. [...]
>
> Le protocole de recueil des données n'est pas toujours très formel, les données sont souvent partielles, mais surtout, quand les résultats sont rapportés dans les communications scientifiques, ils sont épurés de leur contexte et ne souffrent donc aucune critique sur leur interprétation. Enfin, le contexte global du dispositif de formation n'est jamais disponible pour le lecteur étranger à l'expérimentation. Dans ces conditions, il est bien difficile de comparer les méthodes ou d'ouvrir le débat scientifique sur les résultats d'analyse.[22]

Si j'ai choisi cet extrait du rapport, c'est parce qu'il comporte des termes à la polarité clairement négative : « manque de méthodologie », « manque de réplicabilité ou de validité des résultats avancés », « protocole de recueil … pas toujours très formel », « données souvent partielles », « résultats épurés de leur contexte … donc aucune critique sur leur interprétation ». Un

21 http://redac.univ-tlse2.fr/

22 Rapport final du projet Mulce, T. Chanier, version du 17 mars 2011, consultée en ligne le 8 mai 2017 à : http://mulce-doc.univ-bpclermont.fr/IMG/pdf/rapport_fin_de_projet_ANR_Mulce_110317.pdf

inconvénient majeur de l'approche sur corpus est que, paradoxalement, elle se fonde sur des corpus, c'est-à-dire sur des données en quantité trop grande pour être toutes exposées dans les articles qui publient les recherches, à l'inverse des données de l'approche introspective, qui peuvent être présentées *in extenso* à l'appréciation du lecteur scientifique – qui parfois ne se prive pas de contester les astérisques et les conclusions qui dérivent des jugements de grammaticalité et/ou d'acceptabilité. Que l'on valide ou non cette approche, elle a au moins l'avantage de « mettre sur la table » l'ensemble des faits qui soutiennent le raisonnement du chercheur.

J'ai, dans (Jacques, 2016b) [30], mis en évidence l'orientation première de la linguistique de corpus issue de la tradition anglo-saxonne. Son orientation est essentiellement sémasiologique, pour des recherches qui s'attachent aux propriétés des unités lexicales et cherchent à en décrire le sens et la combinatoire. Par exemple, Charolles (2004), ou, plus récemment, Gréa (2015), explicitent le fonctionnement d'unités précises : *sinon*, *parmi*, *entre*, en travaillant à partir de corpus déjà constitués et disponibles sous format électronique (Frantext, Le Monde ou le Monde Diplomatique). Ainsi que je l'ai précédemment mentionné, ce type de recherche présente une facilité indéniable pour le recueil des données : un concordancier donne un accès immédiat à l'ensemble des occurrences présentes dans le (ou les) corpus et même si le chercheur décide d'éliminer certaines occurrences, il est relativement aisé pour d'autres linguistes de retrouver l'ensemble de données à partir desquelles la recherche a été menée. Et ce parce que deux facteurs essentiels autorisent cette possibilité :

1. la recherche porte sur des formes « simples » – même si, pour la forme « entre », une ambigüité catégorielle complique un peu l'affaire ;

2. les corpus utilisés sont accessibles à tous (sous réserve de financement, à savoir abonnement à Frantext et achat des CD-ROM du Monde ou du Monde Diplomatique).

Mais, même si la tradition anglo-saxonne, très imprégnée de préoccupations lexicologiques et lexicographiques (Léon, 2008), met à l'honneur ce type d'études, les linguistes de corpus ne se focalisent pas exclusivement sur des formes. Ils s'intéressent aussi à quantité de phénomènes justiciables plutôt d'une approche onomasiologique, telle que celle qui a été exposée partiellement en 2.2.2. Il est dans ce cas quasiment impossible pour les chercheurs extérieurs d'avoir accès aux données primaires, celles-ci ne se limitant pas aux occurrences d'une forme que n'importe quel concordancier serait à même d'afficher, et le format réduit des publications scientifiques (une vingtaine de pages) n'autorisant généralement pas l'annexion de la totalité des données réellement recueillies et travaillées pour la recherche. Hormis dans le cas des thèses : S. Leroy (2001), par exemple, fournit en annexe de sa thèse la totalité des antonomases sur lesquelles porte sa description – avec en fait l'ensemble du matériel sur lequel et par lequel sa recherche a été menée. Mais l'honnêteté commande de reconnaitre qu'une telle publication détaillée reste l'exception plutôt que la règle.

Pourtant, la publication des corpus de la recherche n'est pas seulement une contribution à la richesse partagée de la discipline, c'est une façon de permettre la vérification, la reproduction, voire la contestation de la recherche. Si par exemple on souhaite « juger sur pièces » les éléments apportés par M. Charolles sur la sémantique de *sinon* dans les articles du Monde Diplomatique, le fait qu'il s'agisse d'un CD-ROM que tout un chacun peut acquérir rend théoriquement possible la reproduction de son extraction de données.

Donc, et je vais revenir là à Corli et à ce que j'ai développé dans la section précédente, une bonne pratique de l'approche sur corpus – et des linguistes de corpus – consisterait à

généraliser la publication des données primaires, au moins avec des accès limités et des outils d'interrogation, comme ceux qui par exemple sont proposés par Frantext ou par Scientext[23], dans les cas où des problèmes de droits se posent.

Et il faut à mon sens aller plus loin. Les propositions qui suivent ne tiennent absolument pas compte de la réalité économique de la recherche, mais des exigences internes d'une démarche scientifique.

Dans tous les types de recherche qui imposent une sélection des données, donc en particulier dans le cas de recherches s'appuyant sur une démarche onomasiologique, la confrontation des jugements opérés sur les données devrait non seulement être mise en œuvre mais aussi publiée avec les analyses des données.

J'ai évoqué dans (Jacques, 2016b) [30] diverses études sur corpus dont l'approche est onomasiologique en ce qu'elles visent à inventorier les formes d'un phénomène linguistique. À l'inverse de l'orientation sémasiologique de la linguistique de corpus de tradition anglo-saxonne, ces études ne partent pas d'une forme pour en décrire les propriétés mais partent d'une signification ou d'un phénomène et cherchent à préciser « comment cela s'exprime ».

Hearst (1992) pour l'anglais puis Borillo (1996) pour le français ont décrit les modalités d'expression de la relation lexicale d'hyperonymie telles qu'elles se manifestent dans des textes « spécialisés » du type encyclopédie ou textes scientifiques. Il s'agissait de produire une liste d'éléments lexicaux et de structures phrastiques en vue d'une utilisation automatique ultérieure. L'objectif pour Hearst était d'enrichir automatiquement ou semi-automatiquement une ressource telle que WordNet de couples hyperonyme-hyponyme. Borillo, pour le français, poursuivait un objectif similaire mais en fournissant une description linguistique bien plus fine des structures analysées. C'est de description linguistique fine qu'il s'agit aussi dans le travail que j'ai déjà mentionné de Rebeyrolle (2000) sur la définition telle qu'elle peut apparaître spontanément dans les textes. C'est encore la description linguistique et la théorisation linguistique qui sont visées par Leroy (2001) dans son travail basé sur un corpus de textes de presse et ambitionnant de « proposer une description et une analyse linguistiques du phénomène de l'antonomase du nom propre » (p. 11). Cette recherche veut aboutir à « une grammaire de l'antonomase ». La grammaire en question se situe dans la lignée des travaux précédemment évoqués, en ce qu'elle se présente sous la forme d'une modélisation-abstraction des énoncés antonomasiques analysés – et qui sont tous, je le rappelle, fournis dans la thèse.

C'est un objectif similaire de description de structures et d'élaboration de « grammaires » destinées à un repérage automatique que poursuit Florez (2014) sur la citation qu'elle appelle « positionnée ». Dans le même esprit que Leroy, elle élabore des patrons de recherche automatique afin de recueillir dans son corpus (thèses et articles scientifiques rendus disponibles par le projet Scientext) les passages dans lesquels les auteurs citent les travaux d'autrui tout en se positionnant par rapport à ces travaux.

De mon côté, une part importante de mon travail de recherche s'inscrit dans cette orientation onomasiologique, comme on le verra dans les chapitres suivants.

Parmi ces études, à ma connaissance, seule Leroy (2004) a testé la reconnaissance spontanée de l'antonomase auprès de 22 « informateurs ». Or, ainsi que je l'ai déjà souligné, dans quasiment tous les cas où des jugements individuels divers sont sollicités, on observe des

23 Je rappelle que Scientext est la base de textes scientifiques développée au Lidilem, elle sera évoquée plus précisément dans les sections 3.2 et 5.3 : http://scientext.msh-alpes.fr/scientext-site/spip.php?article1

divergences. En mesurer et en indiquer l'ampleur me parait donc un temps nécessaire de la recherche.

Et il faut, autant que possible, aller au-delà. Si l'on travaille sur des corpus qui sont eux-mêmes rendus publics, annoter, dans ces corpus mêmes, les données qui ont été sélectionnées pour la recherche hisserait la linguistique de corpus au même niveau d'exigence scientifique que les sciences expérimentales, qui ont bâti leur réputation de sciences « dures » sur les exigences de vérifiabilité et de reproductibilité dont elles se sont dotées au fil de leur développement.

Cette voie, exigeante conceptuellement et techniquement, en ce qu'elle nécessite l'élaboration d'un schéma d'annotation ainsi que sa réalisation concrète, constitue un prolongement logique d'une approche sur corpus et en augmente la portée. Un projet comme Annodis[24] est à cet égard exemplaire : il cumule la constitution d'un corpus informatisé, l'analyse en corpus de phénomènes discursifs (relations rhétoriques, énumérations, chaines topicales...), la production de savoir sur ces phénomènes et la mise à disposition du corpus dans lequel les phénomènes sont annotés, avec le guide d'annotation utilisé. Ainsi, tout chercheur peut reproduire la recherche et/ou juger par lui-même des choix faits sur les données, sans parler de la possibilité de réutiliser sur un autre corpus le schéma d'annotation proposé.

Cependant, pour compléter la réflexion méthodologique que je développe ici, je voudrais tout de même m'attarder quelque peu sur cette dernière proposition, afin de l'éclairer et de la tempérer.

La proposition d'annoter directement dans les corpus les données peut se comprendre de diverses manières. On peut voir cette annotation comme un simple repérage, analogue à un surlignage - ou un soulignage ou toute autre marque consistant à délimiter un segment quelconque dans un texte. Par exemple, dans l'extrait suivant de l'un des corpus sur lesquels j'ai travaillé le phénomène de réduction des termes complexes[25], la forme « ventrale » vaut en fait pour « sangle ventrale » :

(7) Inutile de relâcher exagérément la **ventrale**. Un 40-42 cm, permet une mobilité suffisante. [Vol libre]

Un simple repérage consisterait simplement à entourer les occurrences de cette forme d'un balisage xml qui signalerait que cette forme fait partie des données considérées comme « réduction de terme » :

(8) Inutile de relâcher exagérément la <reduction>**ventrale**</reduction>. Un 40-42 cm, permet une mobilité suffisante.

Ce repérage permettrait de distinguer les cas où « ventrale » est une réduction de terme, des cas où il s'agit d'un adjectif régulier, comme en (9).

(9) Le portage est acceptable, amélioré par de larges bretelles matelassées et une sangle ventrale. [Vol libre]

Mais une autre manière de penser l'annotation ne se satisfait pas de ce simple repérage et intègre au balisage des informations supplémentaires. Ici ces informations pourraient être le terme source de la réduction, le type de réduction, etc. :

(10) Inutile de relâcher exagérément la <reduction type="suppressionTeteNominale" source="sangle ventrale">**ventrale**</reduction>. Un 40-42 cm, permet une mobilité suffisante.

24 http://redac.univ-tlse2.fr/corpus/annodis/

25 Qui sera développé dans le chapitre 3.

Quand on prend en considération l'effort à consentir pour la moindre annotation, on pourrait se dire que cet effort serait mieux rentabilisé en enrichissant, au moment de l'annotation, les occurrences d'informations plutôt que de s'en tenir à un modeste repérage. Mais le gain peut n'être qu'apparent et la richesse n'a pas que des avantages.

Une annotation riche suppose en effet l'élaboration d'un schéma d'annotation, en principe cohérent avec les objectifs de la recherche. Ce schéma est alors plutôt un résultat de la recherche, ou tout au moins un résultat d'une étape liminaire importante de la recherche, qu'un point de départ. Depuis 2007, chaque année, les *Linguistic Annotation* W*orkshops* voient des articles uniquement dédiés à l'élaboration et à la présentation de schémas d'annotation de tel ou tel phénomène (par exemple les phrases génériques dans un corpus, les échanges sur les réseaux sociaux, le sémantisme des prépositions...).

Seconde réserve à une annotation riche, les choix qui sous-tendent les schémas élaborés ne sont pas seulement liés aux questions de recherche des chercheurs impliqués mais aussi à leurs références théoriques. Ainsi, pour rendre compte de la relation dans un texte entre deux phrases adjacentes, peut-on prendre pour guide d'élaboration d'un schéma d'annotation la RST (Rhetorical Structure Theory) (Mann & Thompson, 1988) ou la plus formelle SDRT (Segmented Discourse Representation Theory) (Busquets, Vieu & Asher, 2001) ou encore adapter les règles de progression thématique proposées par Combettes (1988). Le schéma d'annotation est alors en même temps l'outil d'analyse des données. La production des données et le travail sur ces données ne sont ainsi pas dissociés.

Cette façon de faire est intéressante si l'objectif (ou l'un des objectifs) de la recherche est de proposer un schéma d'annotation mais elle est en même temps assez « enfermante » pour les chercheurs extérieurs à la recherche. Ces derniers ne partagent en effet pas nécessairement les cadres théoriques sous-jacents. Ou bien ne sont pas intéressés par une annotation complexe, ou ne parviennent pas à entrer dans sa logique, simplement parce que leurs questions de recherche sont différentes…

En définitive, je voudrais plaider ici pour une séparation entre données et analyse. Pour la lisibilité et la reproductibilité des recherches, les données sur lesquelles le travail est mené devraient idéalement être rendues publiques. La publication des données repose sur, d'un côté, la mise à disposition des corpus d'étude et, d'un autre côté, le repérage des occurrences analysées. Celui-ci, comme je pense l'avoir montré, constitue en fait déjà un premier temps de l'analyse, en ce qu'il engage des décisions de sélection. Pour une claire compréhension des résultats d'une recherche, disposer du corpus d'étude ne suffit donc pas, l'idéal serait d'avoir accès aux données sélectionnées elles-mêmes, en tant que matériau « brut » de l'analyse.

L'analyse produite par la recherche constitue généralement la matière des articles scientifiques, elle représente la dimension plus personnelle de l'apport des chercheurs. Si l'analyse se traduit par une annotation, celle-ci peut être diffusée de façon séparée, afin que l'ensemble des chercheurs puissent au choix bénéficier des analyses effectuées ou aborder les données avec un regard différent. On concilierait de cette façon à la fois le besoin de pouvoir si nécessaire revenir au matériau brut pour le retravailler et la dimension cumulative de la recherche, qui consiste à poursuivre des recherches pour les mener plus loin que l'état de l'art d'un moment donné.

2.3 Synthèse-bilan du chapitre

Dans ce chapitre, j'ai voulu faire le point sur mon positionnement théorique, clarifier mon approche de la linguistique, cerner les implications de choix théoriques et méthodologiques, avancer et développer des propositions méthodologiques. Cette dimension méthodologique et épistémologique est régulièrement abordée dans mon travail de recherche, par des publications (Jacques, 2005a, 2016b ; Jacques & Poibeau, 2010) [12,30] et par des communications à l'occasion de colloques ou de journées d'étude (Jacques, 2001, 2002a, 2012 ; Jacques & Rebeyrolle, 2005). En guise de bilan, je reprendrai ici sous une forme synthétique les points essentiels développés dans le chapitre :

- la perspective claire et affirmée d'une approche sur corpus ;

- une conception de la langue non comme système indépendant clos sur lui-même mais comme partie-prenante du fonctionnement social, cognitif, affectif de l'être humain (je n'aborde toutefois pas toutes ces dimensions dans ma recherche) ;

- l'adoption de principes explicites pour la production des données, la mise en œuvre d'une évaluation de la production de ces données ;

- des analyses autant que possible guidées par le corpus, tout en reconnaissant la place indéniable de la connaissance de la langue à tous les moments de la recherche ;

- la diffusion et la publication non seulement des résultats de la recherche mais aussi de la « matière première » : corpus et données ;

- une distinction claire entre recueil/repérage des données et analyse.

Ainsi que je l'ai mentionné dans le chapitre 1, mon parcours personnel et des références théoriques puisées dans la linguistique de corpus de tradition anglo-saxonne me conduisent à privilégier comme cadre d'analyse l'unité texte. J'ai abordé cette unité à deux niveaux, en m'intéressant à certaines unités lexicales et aux effets de leur mise en texte, et, de façon converse, en regardant le texte du point de vue de sa structuration.

Les deux chapitres suivants synthétisent et articulent mes travaux à ces deux niveaux.

Chapitre 3 - Unités lexicales en texte

Dans ce chapitre, je retracerai le chemin qui m'a conduite des travaux sur les unités terminologiques appréhendées dans leur « habitat naturel », c'est-à-dire les textes de spécialité (Collet, 2000), aux textes scientifiques et à un autre type d'unités qui le peuple : le lexique scientifique transdisciplinaire.

3.1 Des connaissances aux discours spécialisées

Comme je le rapportais dans le chapitre 1, l'origine de mon questionnement sur la langue est double : d'un côté l'énigme que constituait le caractère résistant de la langue écrite, dont l'apprentissage ne se réalisait pas avec le même « naturel » que celui de la langue orale, d'un autre côté la connaissance et la maitrise de ces unités concrètes de la langue que sont les mots.

Les mots, ou plutôt les unités lexicales, sont la partie visible de la spécialisation d'une langue aux fins de servir un ou des domaines de la connaissance. Dès lors que sont élaborés des concepts, des objets, concrets ou abstraits, que sont mises en place des interactions avec ces objets, des systèmes avec ces concepts, de nouvelles unités de toutes natures sont forgées – parfois par transformation sémantique d'unités existantes. La spécialisation de la langue est étroitement connectée à l'élaboration, la transmission et la négociation de savoirs.

Les travaux que je vais synthétiser dans cette première partie du chapitre ont exploré le fonctionnement de ces unités lexicales non pas dans leur relation à leurs référents hors la langue, mais comme éléments des discours par lesquels se construisent et se diffusent des connaissances spécialisées [2, 3, 5, 6, 7, 15].

3.1.1 Termes et textes : situation du problème

Les implications de cette approche que je qualifierai de discursive sont multiples. Il s'agit de concevoir les termes, ces unités lexicales spécifiques d'un domaine de la connaissance ou de l'activité humaine, comme travaillés par le / les discours qui circule-nt dans le domaine en question et non comme de simples étiquettes pour des construits qui se formeraient hors la langue, ce qui serait une « conception mécanique du couplage concept/mot [qui ne prend] pas en compte la complexité des phénomènes langagiers » (Slodzian, 1995 : 10). Les phénomènes langagiers en question tiennent, dans les textes élaborés dans le domaine, à une dynamique liée à la mise en texte qui influe sur ces termes. Ainsi que je l'explicite plus loin, mon travail a consisté à mettre en lumière une part de cette dynamique, prenant ce faisant très au sérieux les conclusions de M. Slodzian, par lesquelles elle préconisait d'intégrer à l'étude terminologique « le syntagmatique, c'est-à-dire les termes en fonctionnement dans les textes » (Slodzian, 1995 : 17). On peut souligner sur ce point la convergence avec les diverses approches linguistiques qui considèrent que la description du lexique doit rendre compte de la combinatoire des unités en tant que participant à la construction du sens, par exemple la Théorie Sens-Texte d'I. Mel'cuk (1997) ou le Lexique-Grammaire de M. Gross (1975).

Une autre implication découle de cette intégration du syntagmatique : la construction de la terminologie d'un domaine, c'est-à-dire l'inventaire des termes et la description de leur système de relations, doit prendre pour substrat les textes produits dans le domaine, donc être textuelle. La terminologie textuelle, prônée par Bourigault et Slodzian (1999), est la marque du groupe TIA, « Terminologie et Intelligence Artificielle », qui rassembla en 1995 des chercheurs d'horizons variés, linguistes, informaticiens, linguistes informaticiens et informaticiens

linguistes, autour d'abord de l'élaboration des Bases de Connaissances Terminologiques (BCT), puis par la suite, autour des questions qui se posent à l'occasion de l'élaboration conjointe de terminologies et d'ontologies. Les BCT avaient comme objectif de dépasser les limitations des systèmes à base de connaissances de l'intelligence artificielle en offrant une prise conjointe sur la dimension linguistique et sur la dimension conceptuelle d'un domaine de la connaissance. L'intelligence artificielle, nous dit Otman (1995) dans l'introduction du numéro de la Banque des Mots qui collecte les textes de la rencontre fondatrice de TIA (Terminologie et Intelligence Artificielle), se pose alors la question de la connaissance et du modèle à construire. La modélisation d'un domaine que proposent les systèmes à base de connaissances trouve ses limites dans le fait d'être bâtie sur la représentation que livrent quelques experts interviewés, experts de la discipline mais pas toujours conscients de leurs propres usages des concepts et des termes. De ce fait, le produit obtenu s'avère être un système parfois déconnecté des réalités du champ et parfois même reflet d'une certaine vision de la discipline, non nécessairement consensuelle.

Le groupe TIA (Terminologie et Intelligence Artificielle), qui défendait ces nouvelles conceptions, était bien enraciné à Toulouse puisqu'en étaient membres Anne Condamines et Didier Bourigault, du laboratoire de recherche ERSS (Équipe de Recherches en Syntaxe et Sémantique), et Nathalie Aussenac-Gilles, de l'IRIT (Institut de Recherche en Informatique de Toulouse). C'est au milieu de cette effervescence intellectuelle stimulante qu'un projet piloté par A. Condamines m'a permis de mener un travail linguistique fondé sur des données attestées, collectées dans le « monde réel » et produites dans un cadre professionnel pour une vraie utilisation de travail. Si j'insiste ici sur l'authenticité de ces données, c'est parce qu'elle a consolidé mon orientation initiale[26] de travail sur la langue « tout venant », non littéraire, ainsi que je l'ai explicité dans le chapitre 1, et qu'elle m'a plongée de plain-pied dans la problématique d'une description linguistique guidée par des objectifs applicatifs, problématique sur laquelle je reviendrai dans la partie II.

Le projet a fait collaborer l'ERSS et l'IRIT pour répondre à une demande de quatre acteurs de la Gestion des Déplacements sur l'agglomération toulousaine : la mairie de Toulouse, la DDE, la SEMVAT (qui était l'acteur des transports en commun : à l'époque bus), le SMTC (autorité organisatrice des transports). Il s'agissait pour ces acteurs d'anticiper la mise en place d'un système commun de gestion en harmonisant leurs terminologies, avec l'hypothèse que cette harmonisation leur permettrait d'éviter les malentendus et incompréhensions liés à des langages différents et de forger un système conceptuel et terminologique commun pour leur collaboration dans la gestion des déplacements.

Le travail auquel j'ai contribué consistait à extraire des textes fournis par ces acteurs institutionnels les unités terminologiques pertinentes, à repérer leurs relations, à construire le système conceptuel sous-jacent en exploitant l'environnement textuel de chaque terme et ce, afin de peupler une BCT (Base de Connaissances Terminologiques) développée par N. Aussenac-Gilles à l'IRIT. Cette exploration du corpus, que je désignerai ici sous le nom de Déplacements, a été l'occasion d'interroger le rapport entre langue et connaissance, j'y reviendrai dans la section 3.1.2.1. Elle a surtout constitué une plongée dans les textes spécialisés et une découverte de la malléabilité syntaxique et sémantique des unités dès lors qu'elles sont immergées dans les discours spécialisés. Deux exemples montreront cette malléabilité.

26 Mon mémoire de Maitrise de Sciences du Langage avait porté sur la description partielle d'une interaction enregistrée dans le train que je prenais quotidiennement.

J'ai pour le premier une tendresse particulière car il illustre à merveille la combinaison de glissement sémantique et d'altération formelle mises au service de la conceptualisation dans un domaine.

Il s'agit de l'utilisation très particulière de la forme « carrefour » dans les textes de ces acteurs. Dans son sens propre, un carrefour est un « Lieu relativement large (par opposition au simple croisement) où se rencontrent plusieurs routes, chemins ou rues venant de directions contraires. » (définition fournie par le TLFi[27]). Ce sens est activé dans :

(11) C'est l'exemple de l'aménagement de deux giratoires aux **carrefours** de la RN20 et de la RD4 qui, réalisé en 1992, a résorbé cette zone accidentogène tout en améliorant sensiblement l'écoulement des différents trafics. [Déplacements]

Mais un énoncé tel que (12) manifeste un premier écart sémantique :

(12) Les exploitants de la Police Nationale doivent notamment avoir la possibilité [...] de mettre un **carrefour** ou un ensemble de **carrefours** au clignotant [Déplacements]

Notre connaissance du monde ordinaire rend cet énoncé pour le moins insolite : en principe, un lieu ne peut pas clignoter ! Il n'est pas besoin d'être un sémanticien très fin pour saisir qu'ici, ce que désigne *carrefour* n'est plus l'intersection mais une partie de celle-ci, le feu de signalisation, et ce par une synecdoque généralisante (le tout désigne la partie).

C'est une sorte encore différente de transformation qui produit (13) :

(13) Bouton permutation : Définit si le **carrefour** autorise la permutation de phase dans le cas de la micro-régulation "appel prioritaire". [Déplacements]

Aucun doute n'est permis, *carrefour* ne peut là désigner ni le lieu ni le feu de signalisation. Il est en fait considéré dans ce contexte comme un élément du système informatique qui assure la régulation du trafic par pilotage à distance des feux de carrefour, au moyen d'équipements associés aux feux en vue de les contrôler, équipements fort logiquement nommés *contrôleurs de carrefour*. Un glossaire fourni par la mairie de Toulouse explicite d'ailleurs cette identité (CAPITOUL-2 est le nom du système informatisé de régulation) :

(14) **carrefour** Du point de vue CAPITOUL-2, élément synonyme de contrôleur de **carrefour**. Le **carrefour** est la vision utilisateur du **contrôleur de carrefour**, ce qui permet de s'affranchir des implémentations différentes de ces contrôleurs. [Déplacements]

Cette dernière équivalence, *carrefour = contrôleur de carrefour*, repose tout à la fois sur une synecdoque généralisante du même type que celle qui construit l'équivalence *carrefour = feu de carrefour* et sur un effacement formel de la tête du terme complexe *contrôleur de carrefour*, effacement qui constitue le second exemple de malléabilité que je vais développer plus loin.

Mais avant ce second exemple, un petit bilan sur la forme *carrefour.* Dans les textes de ce corpus, elle recouvre donc trois « significations » ou plutôt trois opérations de référence : vers un lieu, vers un élément implanté sur le lieu, le *feu de carrefour*, vers un constituant du système informatique de régulation, le *contrôleur de carrefour*. Une telle polysignifiance entre en contradiction avec les théories terminologiques bâties sur la conception « nomenclaturiste » du lexique, pierre angulaire de la « Vienna General Theory of Terminology » de Wüster (Slodzian, 1993), théories selon lesquelles le terme désigne de façon univoque un concept, et dont les

27 Trésor de la Langue Française Informatisé, accessible en ligne : http://atilf.atilf.fr/tlf.htm

chercheurs proches de TIA voudraient dégager la terminologie (par ex. Rastier, 1995). Cette polysignifiance a été pour moi le point d'accroche pour ce champ de recherches.

Le second exemple, qui présente une similitude partielle avec la faculté acquise par *carrefour* de pouvoir référer au feu de carrefour ou au contrôleur de carrefour, est emblématique des modifications formelles que subissent les termes au fil du texte. Les deux extraits suivants mettent le phénomène en évidence :

(15) **L'équipe de conduite d'opération** est responsable de la cohérence de l'opération et de la compatibilité des équipements. Elle rend compte en permanence au Comité de pilotage des conditions de réalisation de l'opération et se doit de l'informer de tout évènement susceptible de mettre en cause le planning ou le coût des travaux. [Déplacements]

(16) **La conduite d'opération** est responsable de la cohérence de l'opération et de la compatibilité des équipements. Elle rend compte en permanence aux maitres d'ouvrage des conditions de réalisation de l'opération et se doit de l'informer de tout évènement susceptible de mettre en cause le planning ou le coût des travaux. [Déplacements]

La quasi-identité des deux contextes manifeste l'équivalence des deux expressions *équipe de conduite d'opération* et *conduite d'opération*. C'est là ce qu'à la suite de Collet (2000), j'ai appelé réduction de termes complexes. Collet a travaillé sur cette réduction dans l'objectif d'en élaborer une grammaire, c'est-à-dire de décrire les règles qui rendaient ce phénomène de réduction possible (Collet, 2003). J'ai de mon côté adopté une approche résolument textuelle et discursive, comme on le verra dans la section 3.1.2.2.

L'élucidation du phénomène de réduction des termes complexes rencontre diverses questions. Une première question à traiter était celle de la rareté ou au contraire de l'abondance du phénomène : est-il reproduit dans des domaines variés, touche-t-il peu ou beaucoup de termes ? Une deuxième question concernait le déclenchement et les effets de la réduction : qu'est-ce qui fait qu'à un certain moment du texte, une partie d'un terme complexe peut être omise ? Est-ce que cette omission est « bloquée » par une éventuelle ambigüité résultante ? En d'autres termes, peut-on par exemple trouver des énoncés dans lesquels il serait impossible de décider ce à quoi réfère la forme *carrefour* ou la forme *conduite d'opération* ? Je me suis aussi au passage intéressée au versant cognitif de la réduction – quoique 'cognitif' apparaisse ici comme un mot un peu pompeux.

Pour traiter ces questions, un second corpus a été élaboré, dans un domaine et un genre délibérément différents. L'hypothèse était la suivante : si les mêmes phénomènes sont observables en ayant modifié les variables de domaine et de genre, alors on peut considérer que ces phénomènes tiennent à d'autres variables, qu'il faut élucider et décrire. Toutefois, puisque mon propos était de mettre en évidence des aspects du fonctionnement des termes dans leur « habitat naturel », ce second corpus devait être lui aussi un corpus de textes spécialisés, faisant usage de termes.

Comme de nombreux locuteurs, je pratique dans ma vie personnelle plusieurs activités qui construisent des connaissances spécifiques, donc des textes spécialisés, donc des termes. J'ai choisi le Vol Libre, en raison du fait que je savais déjà y trouver des attestations de réduction de termes. Deux magazines spécialisés, *Parapente Mag* et *Vol Libre*, m’ont gracieusement fourni une vingtaine d’articles décrivant les essais de voiles de parapente.

Armée de ces deux corpus que, pour les désigner plus facilement, je nommerai « Déplacements » pour le premier et « Vol Libre » pour le second, j'ai abordé ces recherches avec l'outillage qui était à ma disposition à l'ERSS. Grâce à D. Bourigault, j'ai pu bénéficier d'une extraction automatique des candidats-termes de ces corpus (Bourigault, 1994). J'ai ensuite mené une exploration systématique des textes en utilisant d'abord SATO, un logiciel élaboré par F. Daoust[28], puis Yakwa, un outil développé par L. Tanguy (Rebeyrolle & Tanguy, 2001). J'ai aussi adopté, avant qu'Habert (2009) n'en décrive les avantages, l'usage de bases de données relationnelles pour stocker et enrichir les occurrences étudiées et leur contexte.

Le contexte dont il est ici question est double : le domaine, qui conceptualise ses objets et donc en considère certains plus centraux que d'autres pour ses activités, le texte, qui rend possibles voire impose certaines opérations. Avec le recul que me donne l'exercice actuel, je crois nécessaire de mieux distinguer que je ne l'avais fait dans ces recherches ces deux niveaux. Je vais donc reprendre les analyses menées sous cet éclairage.

3.1.2 Les analyses

3.1.2.1 Modéliser les connaissances d'un domaine

Mon entrée dans l'analyse des textes du corpus, au moins pour le corpus Déplacements, concernait les connaissances telles que ces textes les mobilisaient, donc les manifestaient. L'hypothèse sous-jacente à cette analyse est que les textes d'un domaine, produits par des « acteurs » de ce domaine à destination d'autres acteurs, donnent accès aux objets du domaine, aux relations entre ces objets, aux actions sur ces objets, etc., ce qui au bout du compte constitue « les connaissances du domaine ». Par exemple, dans un extrait tel que (17),

(17) ERATO est un système bicéphale associant, dans le but de gérer le trafic des voies rapides urbaines de l'agglomération toulousaine, les deux maitres d'œuvre d'exploitation que sont la DDE (agissant pour le compte de l'État et, par convention de mise à disposition, du Département de Haute-Garonne) et ASF. [Déplacements]

la formulation même livre plusieurs « connaissances » : l'existence de quelque chose qui s'appelle ERATO et qui est un système, le fait que ce système est bicéphale, que son existence est liée à l'objectif de gérer le trafic des voies rapides urbaines et que le caractère bicéphale tient à une association entre DDE et ASF (Autoroutes du Sud de la France).

J'ai évoqué précédemment les Bases de Connaissances Terminologiques, produits logiciels qui reposent sur cette même conception que les textes d'un domaine « contiennent » les connaissances du domaine. L'un des objectifs de ces BCT est de permettre de structurer et modéliser ces connaissances, en s'appuyant sur une analyse semi-automatique des textes, laquelle permet l'extraction des termes et de leurs relations. À partir de là, est facilitée la construction de ce que l'Ingénierie des Connaissances a nommé « ontologies » (Aussenac-Gilles & Condamines, 2000 ; Bourigault, Aussenac-Gilles & Charlet, 2004 ; Bourigault & Charlet, 2000). La recherche autour des BCT ne revêt pas exactement le même visage selon la perspective selon laquelle on aborde ces textes spécialisés. Le linguiste s'intéressera aux moyens mis à disposition par la langue pour exprimer ces connaissances, le terminologue s'intéressera aux concepts et à leurs relations, l'ingénieur de la connaissance s'intéressera à la modélisation du domaine. Mais au bout du compte, ce qui est visé, c'est une représentation de ce que les textes « disent » sous une forme qui facilite l'accès aux connaissances, la

28 Il existe maintenant (juin 2017) une version en ligne de ce logiciel : http://www.ling.uqam.ca/sato/outils/sato.htm

visualisation des relations et connexions au sein du domaine, ainsi que la compréhension des termes et de leurs significations (la liste n'est pas exhaustive de ce que permettent les BCT).

Pour la collaboration avec les acteurs du déplacement dans l'agglomération toulousaine, l'accent était mis sur l'identification de connaissances partagées ou au contraire discordantes. C'est donc plutôt l'aspect « identification de concepts » et « mise en relation des concepts » qui a prévalu sur la dimension proprement linguistique. S'est alors posée une question d'apparence simple qui a constitué pour moi une excellente entrée dans le mode de réflexion de la recherche en linguistique : comment identifier le fait que derrière une même unité terminologique, des concepts différents sont construits par des locuteurs différents ? Si on quitte la sphère des domaines spécialisés, c'est le repérage même des distinctions de sens dans les textes qui est en jeu : à partir de quel moment déterminer que l'occurrence que l'on a sous les yeux présente un sens différent de celui des autres occurrences ?

La question touche à l'appréciation de distinctions sémantiques entre occurrences et j'ai développé dans le chapitre précédent la constatation que ce genre d'appréciation, comme celles qui mettent en jeu une interprétation, est sujette à variation. La perception de différences de sens n'est pas spontanée et n'est pas identique chez tous les locuteurs. J. Véronis (2001) rapporte à cet égard une expérience révélatrice. Il s'agissait de vérifier précisément si des sujets humains étaient capables de réaliser des tâches d'identification de sens. Une expérience mobilisant six sujets différents, étudiants en linguistique, a consisté à présenter à ces sujets une soixantaine de lignes de concordances d'un même mot en leur posant une question simple : « ce mot a-t-il un seul ou plusieurs sens dans tous ces contextes ? ». Les étudiants devaient répondre à cette même question pour 600 mots différents, de trois natures différentes : adjectifs, noms, verbes. Moins de la moitié des mots, 45%, ont donné lieu à un jugement identique des 6 juges. Décider qu'un mot n'a qu'un seul sens ou en a plusieurs se révèle donc moins trivial qu'il n'y parait.

Et lorsque l'on aborde un domaine spécialisé, pour lequel le sens des textes échappe à l'analyste, la probabilité de ne pas percevoir des différences de sens ou de référence augmente. Il est alors nécessaire de définir des critères formels qui permettent de s'affranchir de l'interprétation – au moins partiellement, car on ne peut suspendre totalement la compréhension de la langue, et donc l'interprétation. La problématique est donc au bout du compte de déterminer quels critères mobiliser pour juger que deux unités différentes ont la même valeur.

La solution qui a été suggérée dans (Jacques & Soubeille, 2000) [2] met en jeu la dimension relationnelle et les propriétés des systèmes terminologiques et conceptuels des domaines spécialisés. En partant des réalisations des termes dans les textes, nous avons construit le système de chacun des acteurs des déplacements, sous forme d'un réseau qui rend compte aussi bien des relations hiérarchiques que de relations fonctionnelles. Par *relations hiérarchiques*, j'entends les relations hyperonyme/hyponyme ou partie-tout ; par *relations fonctionnelles*, j'envisage les relations propres à un domaine de la connaissance (pour une étude fine des types de relations, cf. Grabar & Hamon, 2004 ; voir aussi Aussenac-Gilles & Séguéla, 2001)[29]. L'extrait (18) montre un exemple de relation d'hyperonymie, (19) montre une relation spécifique à ce domaine, que l'on peut formaliser ainsi : <agent><contrôle><objet>.

29 Dans le chapitre 5, section 5.2, je présenterai un travail que j'ai mené sur les relations conceptuelles.

(18) Le CRICR est un service collégial de l'État, composé de représentants du ministère de l'Équipement, de la police nationale et de la gendarmerie nationale. [Déplacements]

(19) Les exploitants de la Mairie de Toulouse [...] doivent notamment pouvoir : piloter un carrefour en mode manuel télécommandé [Déplacements]

Après l'élaboration de ces réseaux correspondant à la modélisation des relations exprimées dans les textes, l'enjeu était de cerner les identités et les différences. Nous avons adopté une position très saussurienne en considérant comme ayant la même valeur des unités qui entretiendraient les mêmes relations dans le système. Je vais reprendre l'un des exemples travaillés pour illustrer ce point.

Le terme d'*utilisateur* revient dans tous les textes, comment savoir (et en rendre compte) s'il a la même valeur pour tous les acteurs ? Notre réponse est de prendre en considération les contextes dans lesquels il apparait et les relations que manifestent ces contextes.

Pour une raison très pratique (je n'ai plus à ma disposition tous les corpus travaillés à l'époque), j'illustrerai les similitudes et les différences à partir des occurrences présentes dans deux corpus, celui de la Mairie de Toulouse et celui de la DDE (Direction Départementale de l'Équipement) de la Haute-Garonne. On pourra se reporter à (Jacques & Soubeille, 2000) [2] pour l'étude complète.

Pour ces deux acteurs, lorsque le terme *utilisateur* apparait dans le contexte d'un système informatisé de gestion des déplacements, il renvoie à l'utilisateur du système, comme on le voit dans (20), qui énumère des fonctions que le système doit proposer à l'utilisateur et dans (21), qui est un extrait du glossaire du système Capitoul-2, qui permet d'interpréter plus aisément (22).

(20) Cette opération [hiérarchisation et classement des fonctions] s'effectue selon les critères relatifs au besoin à préciser en cours d'étude, par exemple : […] * type de fonction **utilisateur** (acquisition, consultation, traitement (temps réel ou différé), commande, aide à la décision, aide à l'action, gestion de la maintenance, ...) [Déplacements - DDE]

(21) **utilisateur** Personne utilisant CAPITOUL-2 [Déplacements - Mairie]

(22) Les modules proposés pour répondre aux différentes activités des **utilisateurs** sont :- Préparation, (activités de paramétrage et de préparation de la régulation).- Supervision, (activités de supervision des états des équipements et de la régulation).- Maintenance, (activités de gestion des actions et des équipes de maintenance).- Consultation, (activités de consultation des paramétrages de la régulation). [...]- Administration, (activités d'administration logicielle du système CAPITOUL-2). [Déplacements - Mairie]

Ces occurrences énumèrent des actions assez similaires des *utilisateurs* et peuvent laisser penser que l'unité a la même valeur dans les deux systèmes. Mais la DDE inscrit *utilisateur* dans une autre taxinomie, celle des usagers non pas du système informatique, mais du réseau routier, comme on le voit dans (23) (pour alléger, je supprime les définitions de chaque catégorie d'utilisateur) :

(23) 1. 1. 4 LES USAGERS D'ERATO
Ce sont donc des automobilistes qui circulent sur le domaine considéré qui peuvent se classer en : **utilisateurs** locaux […] ; **utilisateurs** en transit […] ; **utilisateurs** du Département et de la Région […]. [Déplacements - DDE]

Les textes, et en particulier ce contexte, posent une relation d'hyperonymie entre *usagers* et trois autres unités construites à partir du terme *utilisateur*. Cette différence parait à première vue assez insolite, dans la mesure où souvent les termes plus spécifiques d'un domaine sont construits par adjonction de complément ou d'adjectif, dit autrement de modifieur, à une même tête nominale. Assadi et Bourigault (2000) supposent d'ailleurs que « les modifieurs représentent souvent des spécialisations du syntagme nominal tête en spécifiant un attribut de celui-ci » (p. 246) et basent l'organisation de termes extraits automatiquement à partir de textes sur cette propriété. On peut avoir l'impression qu'alors la terminologie dans ce corpus n'est pas très régulière, car on attendrait des termes construits par adjonction de modifieurs à *usagers*, et effectivement les textes mentionnent aussi – et de façon plus fréquente – des *usagers locaux*, des *usagers en transit ou en trafic d'échange avec l'agglomération*, équivalents aux termes construits sur *utilisateurs*.

Cette exploration permet de statuer sur *utilisateur* et de conclure sur l'équivalence seulement partielle de l'unité pour ces deux acteurs : elle a des valeurs très similaires lorsque les contextes évoquent le système informatique mis en place par chacun mais présente une valeur supplémentaire pour la DDE, où elle est synonyme de *usager*.

Au-delà de l'exemple, l'intérêt de cette recherche est de poser un certain nombre de jalons pour le traitement de textes spécialisés dans des applications qui ne visent pas strictement la description linguistique. Dans l'esprit de ce que j'ai développé dans le chapitre précédent, il s'agit d'assoir des interprétations, des décisions, des modélisations sur des critères formels et reproductibles, et donc de participer à l'élaboration de méthodes, transférables au-delà de la « commande » (puisqu'en l'occurrence c'est de cela qu'il s'agissait) qui les a vu naitre.

Un tel travail permet aussi de contribuer à l'élucidation des rapports entre langue et connaissance en mettant en évidence le fait que la langue, plus précisément les textes, ne servent pas à désigner un référent mais en construisent une conceptualisation, singulière pour chaque groupe de locuteurs.

Un exemple en est donné par la mention dans ces textes des acteurs institutionnels ou des entreprises qui sont, en principe, les mêmes pour tous. Je reprends ici la schématisation rendant compte de la conceptualisation de *ASF* (Autoroutes du Sud de la France) autour duquel la taxinomie n'est pas la même dans tous les textes. Ces schémas (figure 1) mettent en évidence la relation d'ASF à un hyperonyme différent pour ces trois groupes : pour la SEMVAT, la société ASF est vue comme faisant partie du groupe de pilotage du système de gestion des déplacements, pour la Mairie, elle est un gestionnaire de déplacements et pour la DDE, un gestionnaire du réseau routier.

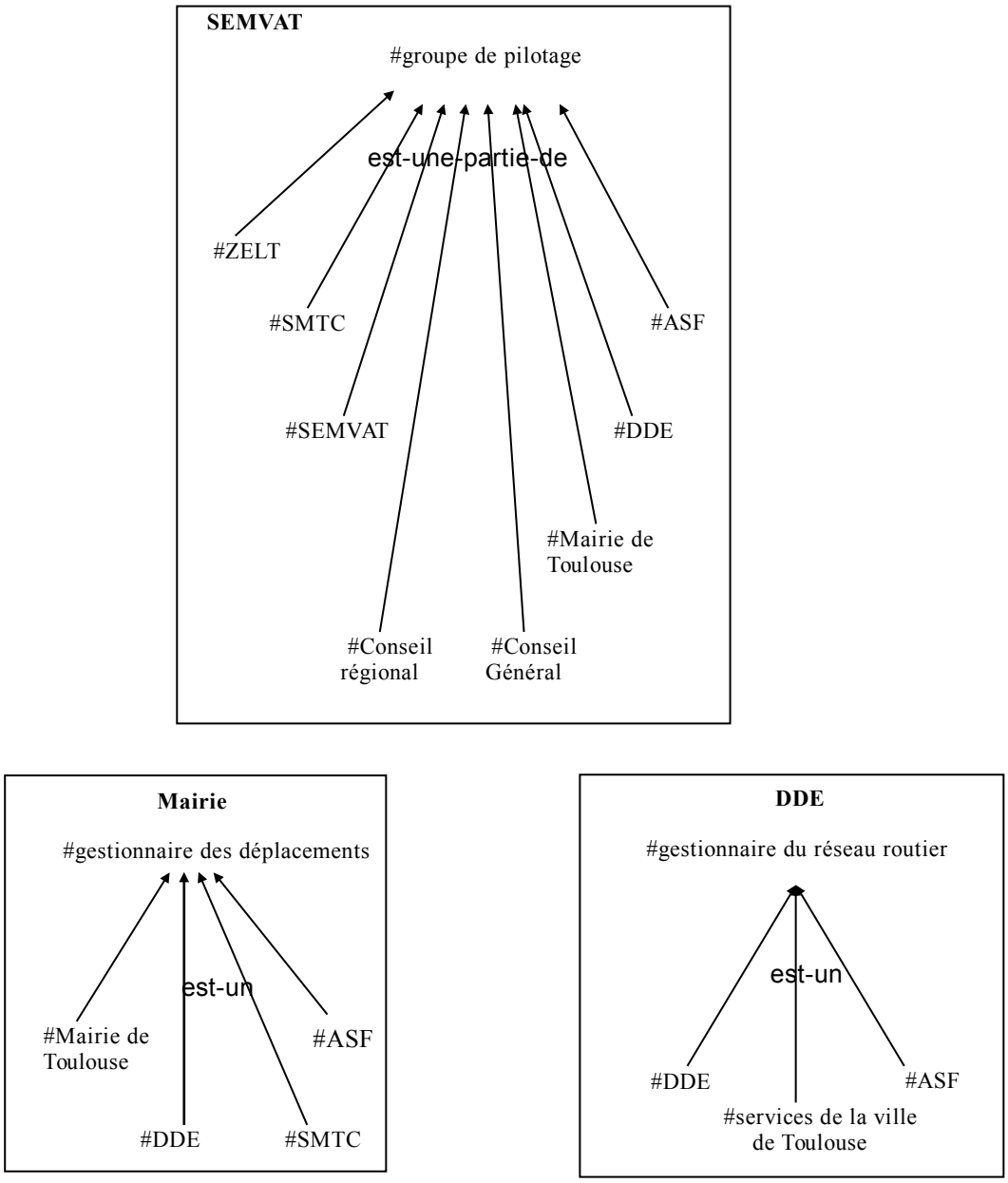


Figure 1 : *Schématisation des relations autour de ASF*

Des différences de cette sorte dans la « mise en texte » de termes identiques, désignant le même référent, sont récurrentes : autre exemple, le *conducteur de bus* n'est pas associé aux mêmes contextes pour les deux acteurs qui l'évoquent dans leurs textes : pour l'un, son interaction avec les passagers du bus est mise en avant, pour un autre c'est son implication dans la circulation des informations à propos du trafic qui est au premier plan.

J'ai interprété ces différences comme témoignant d'une conceptualisation spécifique du référent. Le référent est identique pour tous, c'est la même société ou le même individu, mais les textes le mettent en scène et l'éclairent selon des logiques différentes, ce qui corrobore l'idée que « pour la dénomination spécialisée, il n'existe pas de lien direct entre le référent et le terme mais que cette relation passe par une activité catégorisante. (...) cette activité catégorisante est liée aux spécialités. » (Bouveret, 1997).

Au bout du compte, les résultats d'une telle étude plaident en faveur d'une conception du sens comme non totalement déterminé hors emploi et au contraire fixé par le contexte. Ils montrent aussi et de ce fait la pertinence d'un examen des termes d'un domaine au sein des textes, non seulement parce que ceux-ci sont des « réservoirs » de renseignements indispensables à l'établissement de la terminologie et du système conceptuel d'un domaine (L'Homme, 2005 ; Collet, 2009), mais surtout parce que la « mise en texte » comporte une dimension dynamique, à l'égard du sens – nous venons de le voir – et à l'égard de la forme de certaines de ces unités, aspect que je précise à présent.

3.1.2.2 Identifier et comprendre les phénomènes textuels

Dans mon exploration des transformations que les textes spécialisés opèrent sur les unités lexicales, j'ai été frappée par un phénomène qui avait jusqu'alors été regardé du point de vue de la création lexicale (par exemple Guilbert, 1975) et qui consiste à éliminer une partie à priori très informative d'une unité lexicale complexe pour la raccourcir et la simplifier. Par exemple, dans le contexte des télécommunications, *onde porteuse* devient *porteuse*, et plus généralement, *téléphone portable* ou *ordinateur portable* deviennent *portable*. Comment le nom, qui constitue la tête du syntagme et apporte une information de prime abord essentielle, peut-il ainsi disparaitre ? Plusieurs de mes travaux sont consacrés à l'élucidation de ces mécanismes de réduction (Jacques, 2000, 2002b, 2003a, 2003b, 2006) [3, 5, 6, 7, 15]. Ils se fondent sur la constatation qu'avant d'intégrer le lexique, même spécialisé, les transformations sont d'abord discursives : elles naissent dans les textes, engendrées par des circonstances communicationnelles et/ou des besoins d'expression, puis éventuellement se diffusent par l'usage.

C'est donc sur « mes » deux corpus spécialisés (Déplacements et Vol Libre, évoqués page 54) que j'ai traqué ces modifications, armée de trois principes :

- la réduction devait conserver le sens initial du syntagme source – ce qui s'appréciait par la substituabilité des deux formes ;

- la réduction devait être récurrente et présenter un nombre d'occurrences suffisant pour l'analyse ;

- la réduction devait concerner des termes du domaine.

Ce dernier principe donnait tout son sens à l'entreprise car il me permettait d'ajouter ma contribution aux études soucieuses de réinstaurer les termes dans leur dimension lexicale et textuelle – c'est-à-dire les études qui ne conçoivent pas le terme uniquement comme une étiquette de concept.

Sans retracer tout le cheminement de cette recherche, j'en synthétiserai ici les apports principaux et en montrerai les articulations avec mes autres travaux.

Pour préciser ce dont il est question exactement, j'ai donné dans la partie introductive de ce chapitre deux exemples de la réduction la plus stimulante à étudier, celle qui consiste à supprimer la tête d'un syntagme : de *feux de carrefour* ou *contrôleur de carrefour* à *carrefour* ou de *équipe de conduite d'opération* à *la conduite d'opération*. Dans le corpus Vol Libre, cette réduction réalise la suppression plus classique du nom dans un syntagme Nom-Adjectif : de *ascendance thermique* à *thermique* ou de *sangle ventrale* à *ventrale*. L'effacement de l'autre partie du syntagme est aussi observée, elle produit *chantier* à partir de *chantier courant* aussi bien que de *chantier non courant* (Déplacements) ou *inversion* à partir de *inversion de température* (Vol Libre). Dans les deux cas, ce qui est effacé n'est pas une partie « vide » du

syntagme – rien à voir avec l'économie que font certains domaines des unités grammaticales, par exemple en informatique *accès disque* ou *administrateur système* – c'est au contraire une partie très informative. La réduction pourrait donc sembler aller totalement à l'encontre des besoins communicatifs dans le domaine, qui en général comportent une exigence d'efficacité et de précision. Mais en même temps, la propension des domaines spécialisés à former des termes par composition syntagmatique conduit à des unités certes précises et informatives, mais bien peu maniables dans le discours telles que *système de recueil de données trafic*, *bulletin régional d'alerte météorologique* (Déplacements), *réglage d'origine des commandes*, *sangle ventrale de la sellette* (Vol Libre). Les discours spécialisés mobilisent des moyens divers pour les réduire et les simplifier : les sigles et acronymes, *VRU-LACRA* remplaçant alors *Voies Rapides Urbaines-Liaison Assurant la Continuité du Réseau Autoroutier*, on comprend bien pourquoi, ou ces effacements qu'à la suite de Collet (2000) j'ai appelés 'réduction'.

Les mécanismes à l'origine des deux types de réduction (effacement de la tête ou effacement de l'expansion) sont distincts, le premier se situant dans une logique du discours plus global du domaine, le second dans une logique plus textuelle en ce qu'il met en œuvre le mécanisme de l'anaphore.

1. Réduction par effacement de la tête

Comment les locuteurs d'un domaine peuvent-ils arriver à omettre le nom qui, dans un syntagme terminologique, fournit en général l'hyperonyme pour toute une série de termes ? La suppression de la tête n'est pas une opération courante, elle touche 16 termes complexes sur 99 dans Déplacements et 13 termes sur 40 dans Vol Libre – à noter : l'étude ne portait pas sur la totalité des termes des deux domaines, mais sur ceux qui se voyaient réduits et en même temps présentaient suffisamment d'occurrences pour l'analyse. Le nombre de termes concernés n'est pas important mais le nombre d'occurrences est assez considérable : pour Déplacements, ces 16 termes présentent 639 occurrences de la forme pleine contre 988 occurrences de la forme réduite ; pour Vol Libre, 73 occurrences de la forme pleine contre 790 occurrences de la forme réduite. Les tableaux 2 et 3 indiquent ces termes et leur réduction pour chaque corpus. Pour le terme réduit, si le déterminant est indifférent, il est noté « Dét. », sinon, le déterminant le plus fréquent est mentionné.

On peut voir que, dans les deux corpus, l'effacement concerne des séries de co-hyponymes : les *équipes*, les *systèmes*, les *usagers* dans Déplacements, les *ascendances*, les *élévateurs*, les *fermetures* dans Vol Libre. À partir du moment où la tête se trouve omise pour un élément d'une série, l'existence même de la série facilite l'effacement pour toute la série – c'est particulièrement vrai dans le corpus Vol Libre.

Terme complexe	*Nbre occ*	*Terme réduit*	*Nbre occ*
contrôleur de carrefour	84	Dét. carrefour	128
feu de carrefour	3	Dét. carrefour	7
Direction Départementale de l'Équipement	275	l'Équipement	6
équipe de Conduite d'Opération	27	la Conduite d'Opération	16
équipe de Développement	5	le Développement	11
équipe de Fonctionnement	10	le Fonctionnement	1
équipement (de/du) terrain	55	le terrain	4
ingénieur conducteur d'opération	4	le conducteur d'opération	2
serveur Allo-Trafic	5	Allo-Trafic	3
subdivision des VRU-LACRA	18	la VRU-LACRA	4
système (expert) CLAIRE	22	CLAIRE	25
système Capitoul-2	28	Capitoul-2	358
système ERATO	88	ERATO	404
usager en trafic d'échange	4	le trafic d'échange	7
usager en (trafic de) transit	5	le trafic de transit	8
usager (en trafic) local	6	le trafic local	4

Tableau 2 : *Liste des termes donnant lieu à effacement de la tête avec le terme réduit correspondant dans le corpus Déplacements*

Terme complexe	*Nbre occ*	*Terme réduit*	*Nbre occ*
aile intermédiaire	5	une intermédiaire	7
ascendance dynamique	1	le dynamique	5
ascendance thermique	14	le thermique	119
départ en négatif	2	le négatif	1
élévateur A	10	les A	77
élévateur B	2	les B	77
élévateur C	1	les C	30
élévateur D		les D	22
étape finale	1	la finale	5
fermeture asymétrique	9	une asymétrique	2
fermeture frontale	4	une frontale	7
phase parachutale	23	la parachutale	9
sangle ventrale	1	la ventrale	34

Tableau 3 : *Liste des termes donnant lieu à effacement de la tête avec le terme réduit correspondant dans le corpus Vol Libre*

La description de cet effacement de la tête du terme complexe s'appuie sur sa réalisation dans les textes et sur un aspect plus cognitif. En effet, la réduction réunit diverses conditions :

- la tête – l'hyperonyme – entretient un lien étroit avec la partie restante,
- le référent désigné par le terme est suffisamment saillant dans le discours pour être facilement « récupéré » par le lecteur des textes,
- la relation entre le terme réduit et sa source complète reste donc vivace.

Cette réduction procède ainsi d'un double mouvement : elle survient dans des configurations textuelles qui établissent la relation entre le terme réduit et le terme complexe source, elle joue sur une « récupérabilité » du référent même à travers une désignation qui escamote la partie exprimant l'appartenance générique. Voyons maintenant certaines configurations avant d'expliciter ce qui permet au terme réduit de fonctionner dans le discours sans déperdition d'information.

Plusieurs sortes de contextes permettent de relier terme complexe et réduction, qui établissent cette relation de façon plus ou moins explicite. Le plus explicite juxtapose dans le même énoncé, parfois définitoire, le terme réduit et le terme source, c'était le cas de l'exemple (14), page 55, qui posait *carrefour* comme synonyme de *contrôleur de carrefour*. Dans la même veine, le corpus Vol Libre offre lui aussi une définition en (24) et, en (25), une reformulation du terme réduit *les B* dans une parenthèse explicative qui a aussi pour effet de rendre le terme complexe source transparent :

(24) Quelques rappels pour les débutants :
On appelle **une ascendance thermique**, ou **un "thermique"** tout court (ou encore une " pompe " dans le jargon vélivole et libériste), une masse d'air chaud qui s'élève plus ou moins verticalement dans une atmosphère relativement plus froide. [Vol Libre]

(25) **Les B** (**élévateurs siglés "B"**) tirés de 20 cm conduisent à une Vz de-8 m/s avec beaucoup de stabilité et peu de mouvement dans l'envergure.[30] [Vol Libre]

Du côté moins explicite, se trouvent des contextes qui d'abord utilisent le terme complexe puis, un peu plus loin, le terme réduit. En (26), extrait du corpus Déplacements, duquel je n'ai rien supprimé pour laisser voir l'enchainement textuel, est ainsi d'abord utilisé le terme *l'équipe de conduite d'opération ERATO* puis, dans la section suivante (qui est en fait l'article suivant d'une convention administrative précisant les attributions des partenaires de la gestion des déplacements), le terme réduit *la conduite d'opération*.

(26) Les cahiers des charges des équipements communs et les spécifications fonctionnelles minimales des équipements propres sont définis, après études, par **l'équipe de conduite d'opération ERATO** visée à l'article 13 de la présente convention et approuvés par le Comité de pilotage visé à l'article 12 de la même convention.
5 ARTICLE 5 COUT GLOBAL PREVISIONNEL
Le coût global prévisionnel du système ERATO figure dans la décision ministérielle de prise en considération annexée. **La conduite d'opération** rend compte en permanence au Comité de pilotage de l'évolution prévisible du coût global de réalisation dudit système. [Déplacements]

On peut se douter, au vu des nombres d'occurrences indiqués dans les tableaux 2 et 3, que tous les termes réduits ne sont pas ainsi glissés dans le contexte immédiat de leur terme source, sans

30 Traduction : si on tire de 20 cm sur ces élévateurs B, la voile descend à 8 mètres par seconde sans se tortiller dans tous les sens (ce qui est très appréciable pour le pilote).

quoi, le nombre d'occurrences des deux variantes de termes serait équivalent. On constate au contraire que, particulièrement dans le corpus Vol Libre, mais aussi dans le corpus Déplacements, l'emploi du terme réduit est préféré pour nombre de termes. C'est donc que, d'une part, le terme réduit a gagné son indépendance dans le domaine, d'autre part, malgré cette indépendance, la « récupérabilité » du terme source est effective – sinon, les textes seraient peu compréhensibles.

L'extrait (26) ci-dessus me permet d'aborder ce second aspect, illustré aussi par (12), page 55, et (16) page 56, ou par (27), ci-après.

(27) Le premier **thermique**, rencontré dans la première minute de vol, me satellise d'entrée de jeu à 2200 m. [Vol Libre]

Ces divers extraits mettent en évidence, de façons différentes, une certaine discordance entre la signification littérale du terme réduit et son contexte. Ce n'est qu'en interprétant *conduite d'opération* comme une équipe en (26) ou *carrefour* comme désignant les feux de signalisation en (12) que se résout l'anomalie combinatoire surgie de l'association de ces unités avec des verbes qui à priori ne les comptent pas dans leur sphère de sous-catégorisation. Je l'ai déjà mentionné, avec le sens littéral de *carrefour*, on ne voit pas comment on peut arriver à le mettre en clignotement ou lui envoyer des commandes. L'anomalie combinatoire est ici un déclencheur pour la récupération du terme complexe source.

En (27), le fait que *thermique* soit déterminé par un article défini (***le*** *thermique*) et soit là aussi associé à un verbe qui comporte une idée de mouvement et n'a en principe rien à voir avec la chaleur déclenche là encore la mise en relation avec la notion appropriée, c'est-à-dire l'ascendance thermique.

Ces deux mécanismes conjugués enrichissent un schéma de néonymie comme procédé plus général de lexicalisation selon les étapes suivantes :

1. Création d'un terme complexe, compositionnel et transparent
2. Utilisation de ce terme dans les textes
3. Réduction du terme dans le contexte du terme complexe → coexistence dans les textes de la forme réduite et de la forme complète
4. Diminution de l'utilisation de la forme complète
5. Emploi privilégié de la forme réduite qui se lexicalise

J'ai pu mettre en évidence ce cheminement en prenant en compte la dimension diachronique de certains textes du corpus Déplacements. Ils m'ont permis de retracer l'évolution de *système Capitoul-2* à *Capitoul-2* : d'abord une utilisation des deux formes dans les textes, puis une préférence donnée au terme réduit, et enfin une disparition du terme complexe au profit du seul terme réduit, qui est même posé comme dénomination « officielle » en début d'un document qui explicite le système – voir (28) ci-dessous.

(28) L'objet de ce document est de spécifier les exigences techniques concernant le renouvellement d'une partie du système informatique du PC CAPITOUL qui est désigné dans la suite du document par le terme CAPITOUL-2. [Déplacements]

Le corpus Vol Libre offre une vue plus achevée de cette évolution, dans la mesure où divers textes ne mentionnent même plus le terme source, pariant sur son implantation dans le discours du domaine. De ce fait, d'un terme complexe originellement transparent puisqu'associant un hyperonyme et un modifieur qui assure la partie spécifique, on est passé à un terme plus simple

mais plus opaque, requérant et en même temps instaurant la construction des connaissances spécialisées du domaine.

Ce premier effacement offre donc une vue sur la façon dont les textes spécialisés peuvent « travailler » les termes d'un domaine. Sachant que ces textes sont intimement liés aux connaissances construites dans le domaine, les transformations que le fil du discours opère sont aussi des transformations au niveau de ces connaissances : l'accent est mis sur la partie spécifique du terme originel, l'inscription de la notion exprimée par le terme dans un champ générique n'est plus assurée par la forme du terme elle-même mais doit être déduite de l'anomalie combinatoire des contextes dans lesquels le terme réduit est employé.

La familiarité avec le domaine est ainsi requise pour rétablir la partie effacée. On va voir maintenant que cette familiarité intervient aussi pour le second type d'effacement, qui supprime au contraire la partie spécifique et ne laisse que la tête du terme complexe.

2. Réduction par effacement de l'expansion

À première vue, cette réduction participe essentiellement de la dynamique discursive dans la mesure où elle est favorisée par la situation de reprise anaphorique qui relie clairement le terme réduit au terme source, comme en (29), qui voit *chantier courant* être repris simplement par *chantier*.

(29) Le maître d'oeuvre établit et transmet, sous le couvert du gestionnaire de la voirie (lorsqu'il n'est pas lui-même le gestionnaire), la fiche de prévision à la CDES dans les délais suivants :
* **chantier courant** : au plus tard deux semaines avant la date prévue de début du **chantier**. [Déplacements]

Il y a eu suffisamment de travaux sur l'anaphore pour rendre inutile la répétition de conclusions bien mieux travaillées que je ne pourrais le faire, par des spécialistes aussi éminents que G. Kleiber (1986, 1989, 1990a, 1990b, 2001), F. Cornish (1990, 1996, 2000, 2003), M.-J. Reichler-Béguelin (1988, 1995) ou D. Apothéloz (1995a, 1995b ; Apothéloz & Reichler-Béguelin, 1995).

Je voudrais donc ici attirer l'attention sur deux aspects qui à ma connaissance ont été moins travaillés : i. dans la continuité du discours, jusqu'où considérer qu'une forme réduite constitue une « reprise », autrement dit, sur quelle distance un antécédent textuel étend-il sa portée pour considérer qu'une mention ultérieure de la tête seule est une « reprise » ? ii. des occurrences de termes réduits se passent d'antécédent, elles fonctionnent alors sur une forme de saillance dans le domaine et disent quelque chose de cette saillance.

La première question est indirectement évoquée par Schnedecker et Landragin (2014) dans leur examen des recherches autour des chaines de référence :

> Doit-on considérer que les limites d'une chaîne de référence coïncident avec celles de son texte d'occurrence ? C'est la position implicitement adoptée dans les études menées sur des textes courts. Mais elle semble plus difficile à tenir dès lors que l'on s'intéresse à des chaînes plus longues comme celles de roman : celle renvoyant à Etienne Lantier dans Germinal court-elle effectivement sur les quelque 300 - 400 pages du roman ? Une telle position, du fait qu'elle ne se préoccupe pas du coût de traitement cognitif d'une très (très) longue suite d'expressions référentielles, paraît peu réaliste. (Schnedecker & Landragin, 2014 : 6)

Quoique la notion de chaine de référence et celle d'anaphore ne soient pas directement superposables, comme le font remarquer les auteurs, l'analyse d'expressions référentielles telles que les réductions de termes complexes dans un texte long fait surgir cette question de la relation au sein du texte entre le terme complexe et sa forme réduite à la tête. Comme le montrait l'exemple (29), page 67, une répétition à courte distance favorise la suppression de l'expansion, le mécanisme général de l'anaphore garantissant alors la récupération de ce morceau supprimé par mise en relation avec un antécédent textuel. Cependant, cette réduction par effacement de l'expansion ne se limite pas à des contextes dans lesquels le terme complexe vient d'être mentionné à peu de distance. On observe au contraire pour certains termes une réduction systématique même à très grande distance – ce qui rend difficile la reproduction d'un extrait de plusieurs pages – selon le schéma suivant : une mention du terme complexe, parfois dans un des titres de section, j'y reviendrai, une reprise uniquement par la tête du terme à quelques phrases de distance, puis une réapparition régulière de ce terme réduit, parfois plusieurs pages après le « maillon » précédent.

C'est dans ce genre de configuration que se pose vraiment la question de l'étiquetage du phénomène : s'agit-il encore de reprise anaphorique, d'une continuité référentielle par chainage ? Si l'on s'appuie sur les travaux menés sur la mémoire, on disposerait d'une mémoire de travail, métaphoriquement assimilée au cache de l'ordinateur, ne couvrant qu'un tout petit empan au regard d'un texte long de plusieurs (dizaines de) pages : « le cache est limité à deux ou trois phrases, soit, approximativement sept propositions » (Walker, 2000 : 35). Ce qui signifierait que si l'on utilise des formes telles que « *le réseau* » pour évoquer « *le réseau routier national* » ou « *les équipements* » pour évoquer « *les équipements de terrain* » ou encore « *l'ascendance* » pour parler de « *l'ascendance thermique* » à plusieurs paragraphes voire pages de distance de la mention du terme complexe, le mécanisme de compréhension de ces termes réduits n'est plus analogue à celui qui est mis en jeu pour le traitement des anaphores, mais fait intervenir ce qu'Apothéloz a appelé *saillance cognitive* :

> « la saillance cognitive est un paramètre dont la valeur dépend de la fonction de l'objet dans la schématisation. Un objet sera saillant cognitivement si son statut fait de lui un élément central relativement au sens qui se construit. Un tel objet peut donc demeurer saillant même dans des segments de discours où il n'est pas évoqué » (Apothéloz, 1995b : 315)

La notion de saillance cognitive permet de défendre l'idée que la réduction en ce cas tient moins à un lien de type textuel du terme réduit avec un terme complexe, qui peut être considéré comme son antécédent dans une relation anaphorique, qu'à une importance particulière de la notion dénotée dans le domaine – ou dans l'univers de discours construit par le texte. C'est précisément cette position que j'ai défendue dans (Jacques, 2006) [15] en m'appuyant sur une dernière configuration pour la réduction de termes, dans laquelle le terme complexe et son terme réduit ne sont pas coprésents dans le même texte.

En effet, dans les deux corpus, on trouve des utilisations du terme réduit à la tête pour dénoter ce que dénote le terme complexe. Par exemple, dans le corpus Vol Libre, *ascendance* est fréquemment utilisé pour évoquer une ascendance thermique, sans que ce terme apparaisse, dans des contextes comme :

(30) Il y a suffisamment de vivacité dans les quinze premiers degrés d'inclinaison pour transmettre au pilote des informations aérologiques et donc choisir le "bon" sens de rotation pour se placer dans **l'ascendance**. [Vol Libre]

Pour le lectorat averti auquel ce texte est destiné, aucune hésitation n'est possible, on parle ici d'ascendance thermique et non d'ascendance dynamique : il n'y a de rotation que dans les ascendances thermiques. Cependant, le terme n'est pas du tout mentionné dans le texte en question.

M'appuyant sur la judicieuse remarque de D. Apothéloz, cité ci-dessus, j'ai conclu que ce phénomène témoigne du caractère central de la notion dans le domaine. Plus une notion est saillante et plus elle sera accessible à peu de frais au lecteur du document. L'expression qui la dénote n'a pas alors à être très précise, le contexte se chargeant d'apporter les indices d'une désambigüisation, si nécessaire.

Pour conclure sur ce phénomène de réduction des termes, il apparait comme reposant sur des mécanismes à la fois textuels et cognitifs. Il n'est nullement limité à certains domaines ou certains textes puisqu'il est observé et semble se réaliser selon des modalités similaires dans des univers et des textes de genres différents.

Les textes mettent en place des configurations favorables et, parmi les paramètres propices à sa réalisation, on retiendra le caractère saillant de la notion. Celui-ci peut être intrinsèque au domaine, comme *ascendance thermique* dans le vol libre, ou construit par le texte. Par exemple, dans le corpus d'articles scientifiques que je vais évoquer dans la section 3.2, un article de linguistique intitulé « Approche linguistique pour l'analyse syntaxique de corpus »[31] peut écrire sans souci :

(31) La fonction de notre analyseur est d'identifier des relations de dépendances entre mots et d'extraire d'un corpus des syntagmes (verbaux, nominaux, adjectivaux). Le résultat de **l'analyse** se présente sous la forme d'un réseau de dépendance, dans lequel chaque syntagme extrait est relié à sa tête et à son expansion syntaxiques [Termith - linguistique]

Le thème de l'article, annoncé par le titre, et le contexte de l'occurrence font de *l'analyse* une réduction aisément identifiable de *analyse linguistique*, alors même que le terme ne figure pas dans les phrases qui précèdent l'occurrence.

Les articles scientifiques eux-mêmes éliminent sans hésitation les constituants de termes complexes que la mise en discours rend finalement superflus. Pourtant, l'exemple d'*analyse* montre un potentiel d'ambigüité qui pourrait s'avérer redoutable : le mot est ici tête du terme, mais est dans d'autres contextes une unité lexicale essentielle à la construction du discours scientifique.

3.2 Le lexique scientifique transdisciplinaire : une collaboration fructueuse

Le travail que je vais décrire ici est une collaboration, menée dans le cadre du projet ANR TERMITH[32], avec Agnès Tutin en particulier. C'est à ses travaux que j'ai apporté une contribution très modeste, car elle est au Lidilem la véritable spécialiste de ce lexique, le lexique scientifique transdisciplinaire, et j'ai été conduite à m'y intéresser surtout en raison des interactions de ces unités avec les termes d'un domaine [25, 28].

31 Bourigault, D., & Fabre, C. (2001). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire* 25, pp. 131-151.

32 Projet ANR-CONTINT ANR-12-CORD-0029, http://www.agence-nationale-recherche.fr/?Projet=ANR-12-CORD-0029

Le lexique scientifique transdisciplinaire (LST) est défini par A. Tutin comme un lexique du genre scientifique, motivé par les nécessités de la conceptualisation et de la communication scientifiques :

> Le lexique transdisciplinaire ne renvoie pas aux objets scientifiques des domaines de spécialité, mais au discours sur les objets et les procédures scientifiques. Il peut être considéré de ce point de vue comme un lexique métascientifique (cf. la notion d'opérateur métascientifique dans les travaux de Harris et ses collègues (1989) sur les écrits d'immunologie). Mais le lexique transdisciplinaire des écrits scientifiques ne se cantonne pas au lexique métascientifique. Il inclut aussi un ensemble d'éléments du « métadiscours » au sens de Hyland (2005) ou Dahl (2003). Ces marques linguistiques qui ne renvoient pas au sens propositionnel, mais aux interactions au sens large entre un auteur et son destinataire dans une même communauté, et renvoient à des relations internes au discours. (Tutin, 2007 : 6)

P. Drouin en propose à peu de choses près la même caractérisation :

> Ce lexique, essentiel à l'expression de la pensée scientifique, côtoie la terminologie dans le discours. On considère que le LST transcende les domaines de spécialité et présente un noyau lexical commun significatif entre les disciplines. Le lexique scientifique transdisciplinaire n'est pas saillant dans les textes scientifiques dans la mesure où, contrairement à la terminologie, il se rencontre également dans la langue commune. Par contre, il est au cœur même de l'argumentation et de la structuration du discours et de la pensée scientifique. (Drouin, 2007 : 45)

Ce lexique se définit donc par le fait de situer le discours à un niveau « méta », de servir à la gestion des opérations intellectuelles mises en œuvre et des objets travaillés dans la démarche scientifique, de fournir les moyens de l'argumentation et de la structuration du discours. Synthétisées ainsi, les propriétés décrites en dessinent assez clairement les fonctions et les contours, illustrées par les extraits suivants (je mets en gras les éléments qui relèvent du lexique et de la phraséologie scientifiques transdisciplinaires et j'indique la discipline entre crochets).

(32) **L'objet de cet article est de présenter** la **synthèse** et les **caractéristiques** du discours mythifié qui l'a transformé en un être de légende[33]. [Termith - anthropologie]

(33) Nous nous **proposons** d'**appliquer** la méthode du maximum exact sur diverses séries mensuelles de taux de change[34]. [Termith - économie]

Pour autant, en dresser un inventaire précis n'est pas tâche aisée et c'est à cette recherche que j'ai contribué.

La thèse de Sylvain Hatier (2016), que j'ai co-encadrée avec A. Tutin, a précisément poursuivi cet objectif et nous a offert l'occasion de mener un certain nombre de travaux à partir de l'observation en corpus des diverses unités lexicales qui « peuplent » les textes scientifiques. Je commencerai par présenter le corpus, que je nommerai Termith, avant d'expliciter la teneur de ma contribution.

33 Moine, J. (2006). Basil Zaharoff (1849-1936), le « marchand de canons ». *Ethnologie française*, vol. 36,(1), 139-152. doi:10.3917/ethn.061.0139.

34 Lardic S., Mignon V. (1999). Prévision ARFIMA des taux de change : les modélisateurs doivent-ils encore exhorter à la naïveté des prévisions ? A*nnales d'économie et de statistique*, 54, 47-68.

Le Lidilem dispose, depuis le projet ANR Scientext[35] (ANR-06-CORP-0020), d'une base de textes scientifiques couvrant aussi bien des disciplines des SHS (linguistique[36], psychologie, sciences de l'éducation, traitement automatique des langues) que des disciplines de sciences expérimentales (biologie et médecine) et de sciences appliquées (mécanique, électronique). La partie SHS de ce corpus a été étendue dans un premier temps pour les besoins de la thèse de Thi-Tu-Hoai Tran (2014), par aspiration et traitement d'articles de revues disponibles en ligne, afin de porter le nombre de disciplines à 10 (linguistique, psychologie, sciences de l'éducation, traitement automatique des langues, anthropologie, géographie, histoire, sciences de l'information et de la communication, sciences politiques, sociologie) et le nombre total d'articles à 300. Sylvain Hatier a encore enrichi ce corpus en portant le nombre total d'articles à 500, 50 dans chaque discipline. Le corpus résultant représente presque 5 millions de mots.

Ce corpus a été analysé syntaxiquement et les procédures d'extraction automatique puis de regroupement en classes mises en œuvre par Hatier se sont appuyées sur les relations de dépendance syntaxiques. L'ensemble de la démarche a été synthétisée dans (Hatier *et al.*, 2016). J'emprunte à (Hatier, à paraitre) le schéma récapitulatif des traitements appliqués pour les noms du LST (figure 2). On y voit les trois étapes du traitement, numérotées de 1 à 3, les ressources et outils utilisés, dans les formes rectangles sur fond gris, et les applications envisagées pour le lexique typé sémantiquement obtenu à l'étape 3, reliées par des flèches dans la partie droite du schéma.

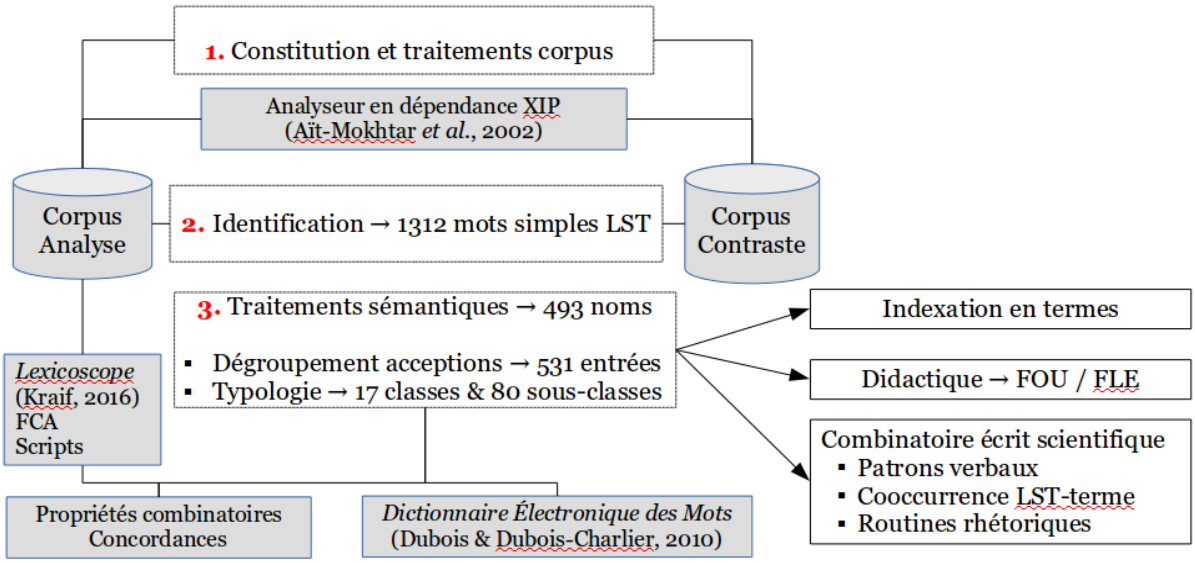


Figure 2 : *Traitements et analyse des noms du LST*

Ma contribution, outre le suivi de la thèse, a porté sur les étapes 1 (la constitution du corpus, comme je l'ai indiqué plus haut) et 2 (identification du LST), et a chemin faisant plus particulièrement concerné le comportement en discours de ces unités lexicales transdisciplinaires : leur cooccurrence avec les termes (Jacquey *et al.*, 2013) [25] et les ambigüités potentielles des formes dans les textes (Jacques & Tutin, 2015)[37]. Comme j'ai

35 http://scientext.msh-alpes.fr/scientext-site/spip.php?article1

36 C'est anecdotique, mais j'avais alors donné ma thèse et un article de recherche pour ce corpus, sans imaginer que je rejoindrais plus tard la stimulante équipe qui l'a constitué.

37 Cette dernière étude n'a pas donné lieu à une publication, faute de temps pour la rédiger dans les délais.

évoqué précédemment la constitution du corpus, je précise maintenant mon apport à ces recherches pour l'identification du LST et pour l'analyse de son fonctionnement dans les textes.

Comme le rappelle (Tutin, 2007, voir plus haut), nombre de chercheurs s'accordent sur le fait que les textes scientifiques font usage de ce lexique particulier, dans des constructions récurrentes qui constituent même une phraséologie typique de l'écriture scientifique (Drouin, 2007 ; Pecman, 2007 ; Siepmann, 2007). Toutefois, passer de cette reconnaissance unanime à un inventaire effectif de ces unités impose un repérage dans les textes et impose donc des décisions en tous points analogues aux jugements que j'évoquais dans la section 2.2.2. Conscients des biais éventuels possibles, nous avons mis en place un protocole comportant une phase de jugements, laquelle a mobilisé 5 juges différents qui se sont prononcés sur les mêmes données[38].

Les données soumises à ces 5 juges étaient les « candidats LST » extraits automatiquement du corpus en fonction de critères de spécificité, de fréquence et de répartition entre les différentes disciplines (pour une description de la procédure d'extraction voir Hatier, 2013, 2016) : au moins 20 occurrences de chaque unité, présence dans au moins 5 disciplines. Pour chaque unité, il fallait juger de son appartenance à l'un des trois types de lexique suivants :

- lexique non spécialisé, dit « de langue générale », c'est-à-dire le lexique utilisé par tout un chacun, dans le sens commun ;
- lexique abstrait général, c'est-à-dire le lexique typique de l'activité de réflexion, présent dans les textes scientifiques mais présent aussi dans tout texte qui élabore une pensée abstraite : « Le lexique abstrait général (LAG), ainsi que nous le nommons, est constitué d'éléments non spécifiques au genre mais sur-représentés dans l'écrit scientifique, tels : *manière, difficulté, enjeu, rôle, inhérent, abstrait, précédent, aborder, confronter,participer, effectivement, strictement*. » (Hatier, 2016 : 16) ;
- lexique scientifique transdisciplinaire.

On notera que ces catégories forment une sorte de continuum, du moins spécialisé vers le plus spécifique, sans aller jusqu'au lexique disciplinaire qu'est la terminologie.

Nous attendions de la confrontation des jugements un éclairage sur la perception que des « experts » de l'écriture scientifique, qui plus est linguistes, pouvaient avoir de la notion 'LST' et de la différenciation de ces types de lexique. Pour aider la décision, les données ont été présentées à travers une interface destinée à faciliter l'appréciation de l'utilisation des unités lexicales, donc de leur valeur en contexte.

La figure 3 montre cette interface, un formulaire de bases de données[39], qui comporte :

- sur la partie gauche supérieure, le nom à traiter et des indications de sa combinatoire ;
- sur la partie gauche inférieure, le jugement à produire – sous forme de case à cocher – ainsi qu'une case pour un éventuel commentaire ;
- sur la partie droite, des exemples de contexte dans différentes disciplines[40].

38 Les 5 juges étaient nos deux collègues de l'ATILF, Evelyne Jacquey et Laurence Kister, et notre équipe projet du Lidilem, Sylvain Hatier, Agnès Tutin et moi-même.

39 On verra au chapitre 4 que j'avais déjà développé dix ans plus tôt une interface d'annotation pour l'analyse des titres de sections, je développerai à cette occasion l'argumentation sur les bénéfices de telles interfaces de type 'formulaire'.

40 On pourra se reporter à (Hatier, 2016 : 83) pour une explicitation des consignes.

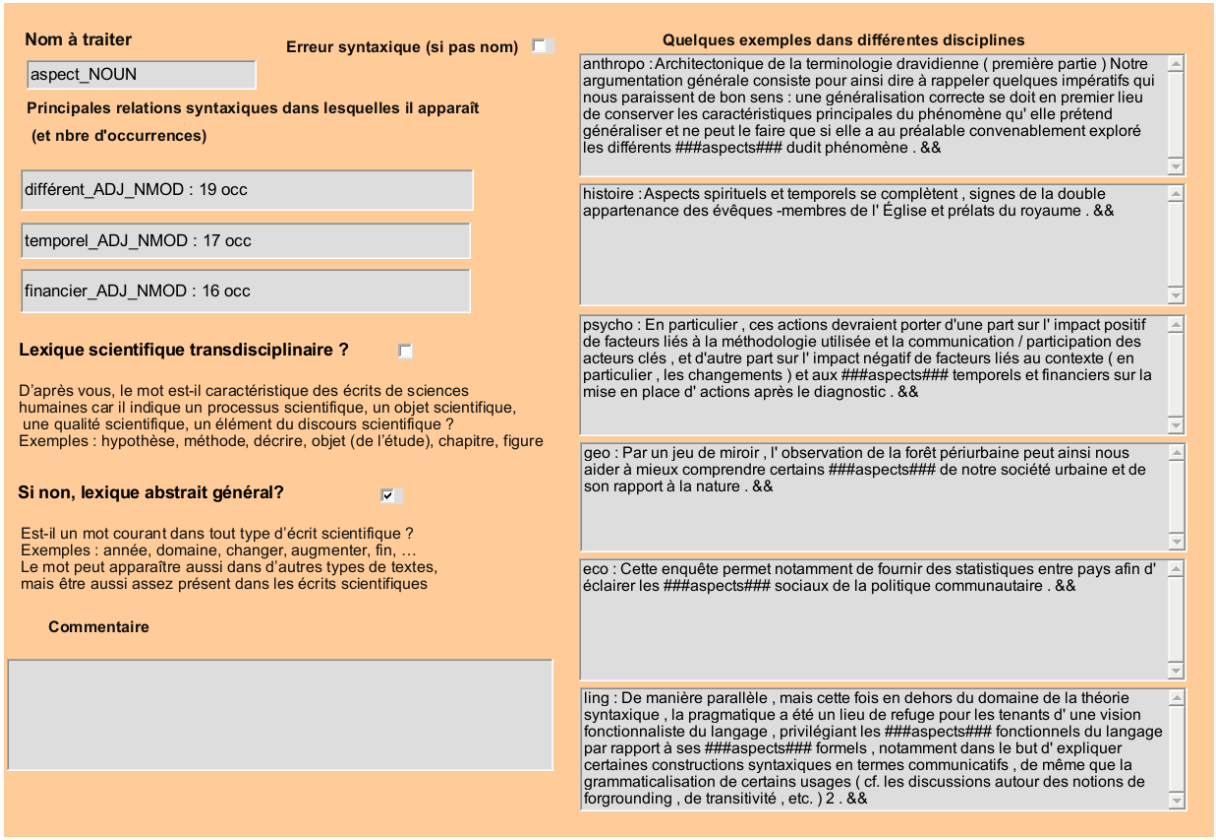


***Figure 3** : Interface d'analyse des candidats-LST*

De même que pour l'analyse des segments procéduraux (cf. section 2.2.2), la convergence des jugements a été appréciée par une mesure statistique, ici le Kappa de Fleiss[41]. Celle-ci met en évidence un certain accord pour la différenciation d'un lexique spécifique et du lexique non spécialisé et une réelle divergence pour la distinction LST / lexique abstrait général (LAG), toutefois variable selon les catégories : convergence modérée pour les noms, forte pour les adjectifs, faible pour les verbes, extrêmement polysémiques.

	LST	LAG	GEN
Noms	0,45 (83 %)	0,251 (64 %)	0,428 (72 %)
Verbes	0,234 (63 %)	0,081 (55 %)	0,337 (82 %)
Adjectifs	0,755 (90 %)	0,615 (83 %)	0,746 (88 %)

***Tableau 4 :** Accords inter-annotateurs – Kappa et pourcentage d'accord* (Hatier, 2016 : 94)

On voit cependant que, toutes classes grammaticales confondues, c'est bien pour le rangement dans le type LAG que la divergence est la plus importante. Ce travail de validation a ainsi débouché sur la décision d'abandonner la distinction entre LST et LAG et de confondre en une seule liste les unités rangées lors de la validation dans l'un ou l'autre type. Ont ainsi été retenus 493 noms, 274 adjectifs et 342 verbes dont au moins une acception relève du LST ou du LAG.

Cette étape de sélection du LST, en nous donnant un aperçu des contextes autour des unités, a permis de cerner une partie de leur fonctionnement discursif sous deux aspects : leur interaction avec les termes des domaines et les potentiellement multiples valeurs des formes.

41 Je reprends ces différents éléments de la thèse de Hatier (2016, chapitre 2).

L'interaction LST-termes a été utilisée dans (Jacques, 2011) [22] avec l'utilisation de patrons définitoires pour l'identification de termes émergents et examinée dans (Jacquey *et al.*, 2013) [25] essentiellement en vue de vérifier la possibilité d'améliorer la détection automatique de termes dans les textes, en faisant l'hypothèse qu'une relation de dépendance syntaxique entre une unité du LST et un terme putatif serait exploitable comme indice du caractère terminologique du candidat-terme. Cette hypothèse s'appuyait sur des contextes tels que (34) ou (35), dans lesquels l'unité du LST, en gras, a sous sa dépendance un terme, en italique.

(34) nous nous attachons à **analyser** et comprendre *les gestes de travail* à partir de leur élaboration, qui permet aux opérateurs de développer des compétences et de se maintenir dans leur activité de travail[42] [Termith - psychologie]

(35) On peut vouloir indiquer par ce **concept** de *langage intérieur* ou bien que la pensée n'est que la forme intériorisée du langage externe, ou bien que la pensée assimilée à une représentation est structurée comme le langage naturel qui n'en est que la manifestation externe[43]. [Termith - linguistique]

L'étude s'est limitée aux textes de linguistique, domaine pour lequel les collègues de l'ATILF avec lesquelles ce travail a été mené disposent de listes de termes et de textes étiquetés en termes. Elle a conclu que l'hypothèse de départ est valide pour certaines classes sémantiques d'unités du LST, par exemple celles qui concernent les processus et contiennent des noms tels que *analyse*, *sélection* (et les verbes correspondant) mais pas pour tout le LST dans son ensemble. Par exemple, si l'on extrait les concordances d'un nom comme *chercheur*, qui fait partie du LST, suivi d'une préposition – afin d'atteindre les contextes dans lesquels *chercheur* est en relation de dépendance avec un autre groupe syntaxique –, on constate qu'aucune des 7 occurrences ne se combine avec un terme (figure 4).

text_id	Contexte gauche	Pivot	Contexte droit
Texte_linguistique_38_XipEmolex	va permettre de remarquer que ce sont les	chercheurs du	Cercle de Vienne qui, " se découvrant des affinités avec
Texte_linguistique_31_XipEmolex	du langage extériorisé, ce qui pousse le	chercheur à	faire l'économie de la question quant à sa spécificité su
Texte_linguistique_31_XipEmolex	analyse des produits de la pensée, les	chercheurs autour de	Külpe analysaient les processus de la pensée qui se dér
Texte_linguistique_24_XipEmolex	de Pour La Science a annoncé que des	chercheurs de	l'INRA ont mis en évidence le fait que les plants de
Texte_linguistique_24_XipEmolex	débarrasser de cette notion de sujet. Les	chercheurs en	logique mathématique utilisent des prédicats à n - argur
Texte_linguistique_56_XipEmolex	Sont -ce des pesanteurs historiques qui conduisent les	chercheurs à	refaire un parcours théorique fondateur qui paraît décon
Texte_linguistique_58_XipEmolex	vendre deux pelles et deux pioches à des	chercheurs de	trésors. (Bataille, L'Arbre de No 'l)

***Figure 4** : Concordances de chercheur + préposition*

En revanche, les 157 occurrences d'une recherche similaire avec *analyse* livrent 82 contextes dans lesquels le GN complément est un terme, tel que « analyse des déictiques », « analyse des verbes préfixés », « analyse de suites conversationnelles », etc. Parmi les occurrences restantes, une petite dizaine évoque d'autres chercheurs, par ex. « l'analyse de Kleiber », « l'analyse de Leeman », « l'analyse de Ducrot », et enfin certaines sont elles-mêmes un terme du domaine, *e.g.* « analyse du discours ».

On se doute que je n'ai pas choisi ce nom au hasard, mais parce qu'il illustre parfaitement le dernier aspect que nous avons traité, et qui concerne la multiplicité des valeurs des formes de ces unités du LST dans les textes. En effet, une occurrence d'*analyse* peut être, selon son contexte :

42 Chassaing, K. (2010). Les « gestuelles » à l'épreuve de l'organisation du travail : du contexte de l'industrie automobile à celui du génie civil. *Le travail humain*, vol. 73,(2), 163-192. doi:10.3917/th.732.0163.

43 Puech C. (2001) Langage intérieur et ontologie linguistique à la fin du XIXe siècle. *Langue française* 132, 26-47.

- une unité de la langue générale, comme par exemple dans « faire l'analyse de la situation » ;
- une unité du LST, extrait (36) ;
- la tête d'un terme complexe, extrait (37) ;
- un terme réduit (par effacement de l'expansion, cf. p. 67), exemple (31), p. 69 ;
- un terme simple hyperonyme d'une série (par ex. *activité* comme hyperonyme de a*ctivité langagière* et *activité métalinguistique*), extrait (38).

(36) Cet article propose une **analyse** des verbes préfixés mettant en jeu le préfixe *sous-*[44]. [Termith - linguistique]

(37) Il est courant en **analyse du discours** de distinguer les aspects situationnels et textuels du discours[45]. [Termith - linguistique]

(38) Toute une partie des postulats, les conventions de la pragmatique, [...] ont tranché et ramené l'activité de langage à une **activité** claire entre des gens qui veulent coopérer pour aboutir à un résultat que le premier voulait avoir en tête et que le second cherchait à dégager[46]. [Termith - linguistique]

Dans (Jacques & Tutin, 2015), nous nous sommes posé les questions suivantes : dans quelle mesure les valeurs envisagées plus haut se réalisent-elles ? Y a-t-il dans un même article scientifique cohabitation de plusieurs valeurs ou une forme d'ambiguïté est-elle évitée par restriction à une seule valeur ? Si plusieurs valeurs apparaissent, dans quelle proportion ? Peut-on dégager des indices, notamment formels, qui permettraient une résolution de cette ambiguïté ? Nous faisions l'hypothèse que l'écrit scientifique tendrait à restreindre l'ambigüité en ne multipliant pas les valeurs possibles ou en fournissant des indices clairs pour l'interprétation en contexte.

Pour l'étude, nous avons sélectionné les noms figurant dans notre liste d'unités du LST et apparaissant aussi comme têtes de termes complexes de linguistique, dans le thesaurus élaboré par l'ATILF. Une vingtaine de noms réunissait ces deux conditions : *acquisition, activité, analyse, article, base, communication, construction, corpus, effet, faculté, groupe, interprétation, modalité, mot, ordre, production, sens, signification, traitement, univers*. Nous avons étiqueté leurs occurrences (1208 occ.) selon leur valeur (l'une des 5 valeurs évoquées plus haut) et avons ensuite observé s'il y avait dans un même texte emploi de valeurs différentes pour une même unité.

De la même manière que je l'avais constaté pour les réductions de termes (Jacques, 2003a) [7], les textes ne se soucient guère de se restreindre à une seule valeur. Pour certaines des unités considérées, l'emploi soit comme tête de terme, soit comme unité du LST, n'altère pas leur sémantisme propre, c'est-à-dire que dans « analyse sémantique », « analyse syntaxique », « analyse du discours » aussi bien que dans « proposer une analyse », le nom *analyse* garde une même signification. Peu importe alors au bout du compte qu'il ait valeur de terme ou de LST, le message n'en souffrira pas car le statut de l'unité ne joue pas sur cette signification,

44 Paillard Denis (2002). Contribution à l'analyse du préfixe *sous-* combiné avec des bases verbales. *Langue française*, n°133, 91-110.

45 Burger Marcel (2000). Scènes d'actions radiophoniques et prises de rôles : débattre, informer, divertir *Revue de Sémantique et de Pragmatique* 7, 139-154.

46 Laurendeau, Paul (1997). Contre la trichotomie Syntaxe/sémantique/pragmatique. *Revue de Sémantique et de Pragmatique* 1, 115-131.

suffisamment sous-spécifiée pour s'accommoder de l'orientation que lui donnera sa combinatoire. Il y a là une situation de *vague* (Fuchs, 1986) dont au final les textes s'accommodent, voire même tirent bénéfice pour s'épargner un effort de précision – Berrendonner (1990) a décrit cette économie en terme de *principe de nonchalance*.

Dans les cas où l'utilisation comme terme entraine un sémantisme différent, par exemple pour *modalité*, la valeur appropriée, soit unité du LST – ex. (39), soit terme du domaine – ex. (40), et donc le sens qui lui correspond, pourra être sélectionnée grâce au contexte autour de l'occurrence, selon l'idée que, dans un énoncé, « chacune [des unités lexicales] n'acquiert son sens qu'en fonction de la présence des autres » (Victorri, 1997 : 47).

(39) V. Egger accomplit, selon des **modalités** que nous développerons, la rupture entre la forme sonore de la pensée et le sens[47]. [Termith - linguistique]

(40) Nous adopterons ici le parti contraire, considérant que l'exclamation présente des points communs avec l'assertion, l'interrogation et l'impératif qui, eux, sont reconnus comme des **modalités** à part entière[48]. [Termith - linguistique]

En fin de compte, dans les textes des SHS et en particulier en linguistique, LST et termes entretiennent des relations diverses, aussi bien syntagmatiques, par la cooccurrence et la dépendance syntaxique, que paradigmatiques, par le fait que certaines unités du LST peuvent entrer dans la construction de termes complexes d'un domaine. Ces relations sont encore loin d'être véritablement explorées, nous en avons à peine effleuré un aspect, la coexistence dans les textes. Il serait sans doute fructueux de poursuivre ce travail sur les versants syntagmatique, paradigmatique, textuel, dans l'ensemble des disciplines de SHS dans la mesure où, en premier aperçu, ces disciplines semblent construire leurs termes et leurs concepts en « empruntant » des unités lexicales qui existent déjà et en les spécialisant, ou en les spécifiant par adjonction d'expansions.

Ce champ de recherches autour du LST a élargi mon approche des textes spécialisés et m'a permis de réinvestir une partie des travaux menés sur la réduction des termes complexes. Dans le chapitre 4, je poursuis l'exposé des recherches menées sur les textes spécialisés, mais en changeant de niveau d'analyse : non plus les unités lexicales mais les objets textuels impliqués dans la structuration des textes longs et qui participent à leur découpage en sections et sous-sections.

3.3 Synthèse-bilan du chapitre

J'ai dans ce chapitre retracé le chemin qui m'a menée de l'étude des termes complexes et de leurs variations dans les textes à celle des unités du lexique scientifique transdisciplinaire. Dans l'analyse des lexiques, je n'ai jamais perdu de vue la dimension textuelle, cherchant à ressaisir et restituer la façon dont le texte, dans son déroulement, ou par les cooccurrences qu'il rassemble en un même empan, phrase, paragraphe ou section, altère les unités, donne des indices pour sélectionner le sens adéquat ou néglige d'éviter les ambigüités potentielles.

J'espère avoir ainsi éclairé d'autres caractéristiques de mes choix en tant que chercheure :

- un intérêt pour l'interaction texte-lexique, avec la perspective d'une action transformante du texte sur le lexique ;

47 Puech Christian (2001). Langage intérieur et ontologie linguistique à la fin du XIXe siècle. *Langue française*, 132, 2001, 26-47.

48 Jacqueline Bacha (2000). Marqueurs exclamatifs et aspect verbal. *Revue de Sémantique et de Pragmatique* 7, 9-28.

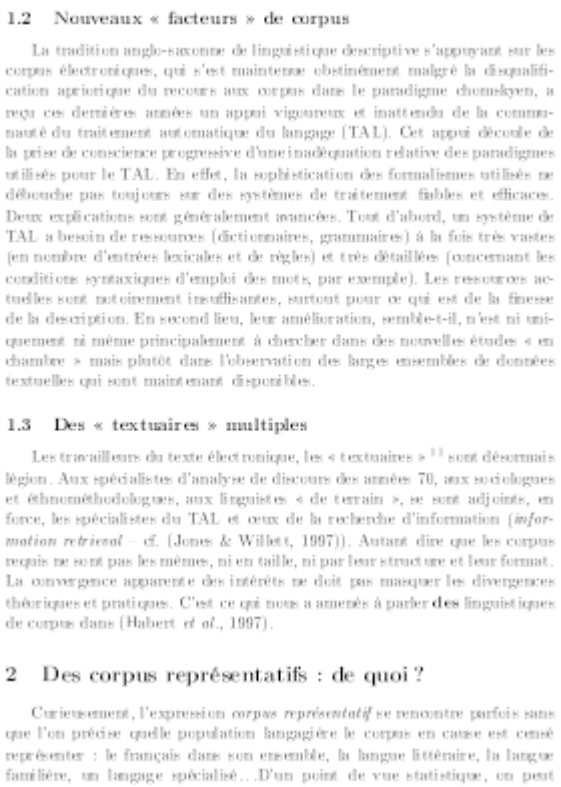
- la mobilisation de procédures d'annotation et plus généralement des méthodes d'analyse évoquées dans le chapitre précédent pour la production des observables et le support des analyses ;

- un intérêt continu et immarcescible pour les textes spécialisés, en ce qu'ils mêlent les particularités linguistiques dont j'ai donné un premier aperçu et des relations fortes avec la construction des connaissances, celle-ci passant par les formes langagières et discursives.

Le chapitre suivant poursuit la description des particularités linguistiques des textes spécialisés mais, comme je l'ai précisé, en regardant le texte au niveau de sa structuration.

Chapitre 4 - Un objet textuel mal identifié : focus sur les titres de section / intertitres

Hormis le corpus Vol Libre, constitué d'articles d'essais de voiles de 2-3 pages, les corpus que j'ai évoqués dans les deux chapitres précédents ont en commun – outre le fait de manifester des usages spécialisés de la langue – de rassembler des textes longs. Le corpus Déplacements est constitué de documents tels que des cahiers des charges, des descriptions d'organisation du travail, des projets de systèmes (de gestion des déplacements humaine et informatique), de 15 à 40-50 pages ; le corpus médical à partir duquel a été travaillée l'expression des procédures (section 2.2.2) est constitué de recommandations médicales d'une dizaine de pages en moyenne ; le corpus Termith est constitué d'articles scientifiques, de 15-20 pages selon les disciplines et les revues… La lecture de ces documents nécessite alors un temps qui excède l'empan mémoriel à court terme – souvenons-nous que les travaux sur le centrage d'attention, cités dans le chapitre précédent, font état d'une mémoire de travail qui couvrirait deux-trois phrases (Walker, 2000). Il semble donc nécessaire que des textes longs disposent de mécanismes qui facilitent leur appréhension sans surcharge cognitive. L'un de ces mécanismes est leur structuration : ces textes sont organisés en sections, sous-sections, paragraphes. Les sections et sous-sections sont généralement – et effectivement dans notre corpus – titrées. Je reproduis dans la figure 5 une page d'un article de B. Habert[49], on y voit trois titres de section : les deux premiers découpent la section 1 de l'article, le dernier introduit la section 2.



1.2 Nouveaux « facteurs » de corpus

La tradition anglo-saxonne de linguistique descriptive s'appuyant sur les corpus électroniques, qui s'est maintenue obstinément malgré la disqualification apriorique du recours aux corpus dans le paradigme chomskyen, a reçu ces dernières années un appui vigoureux et inattendu de la communauté du traitement automatique du langage (TAL). Cet appui découle de la prise de conscience progressive d'une inadéquation relative des paradigmes utilisés pour le TAL. En effet, la sophistication des formalismes utilisés ne débouche pas toujours sur des systèmes de traitement fiables et efficaces. Deux explications sont généralement avancées. Tout d'abord, un système de TAL a besoin de ressources (dictionnaires, grammaires) à la fois très vastes (en nombre d'entrées lexicales et de règles) et très détaillées (concernant les conditions syntaxiques d'emploi des mots, par exemple). Les ressources actuelles sont notoirement insuffisantes, surtout pour ce qui est de la finesse de la description. En second lieu, leur amélioration, semble-t-il, n'est ni uniquement ni même principalement à chercher dans des nouvelles études « en chambre » mais plutôt dans l'observation des larges ensembles de données textuelles qui sont maintenant disponibles.

1.3 Des « textuaires » multiples

Les travailleurs du texte électronique, les « textuaires »[11] sont désormais légion. Aux spécialistes d'analyse de discours des années 70, aux sociologues et ethnométhodologues, aux linguistes « de terrain », se sont adjoints, en force, les spécialistes du TAL et ceux de la recherche d'information (*information retrieval* – cf. (Jones & Willett, 1997)). Autant dire que les corpus requis ne sont pas les mêmes, ni en taille, ni par leur structure et leur format. La convergence apparente des intérêts ne doit pas masquer les divergences théoriques et pratiques. C'est ce qui nous a amenés à parler **des** linguistiques de corpus dans (Habert *et al.*, 1997).

2 Des corpus représentatifs : de quoi ?

Curieusement, l'expression *corpus représentatif* se rencontre parfois sans que l'on précise quelle population langagière le corpus en cause est censé représenter : le français dans son ensemble, la langue littéraire, la langue familière, un langage spécialisé…D'un point de vue statistique, on peut

Figure 5 : *Une page d'article scientifique comportant 3 titres de section*

49 Habert, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? *Cahiers de l'Université de Perpignan*, 31, 11-58.

Je commencerai en explicitant les raisons pour lesquelles je me suis focalisée sur ces titres de section et indiquerai les recherches qui les ont concernés (section 4.1) avant de faire la synthèse de mes propres recherches (sections 4.2 et 4.3) [4, 8, 9, 13, 18, 20, 27, 32, 33].

4.1 Titres et intertitres : objets de l'attention

Pourquoi m'intéresser aux titres ? Pour un contexte tel que (41) :

(41) 2.2.4 **BULLETIN** PREVISIONNEL
Ce bulletin est transmis par télécopieur au CIGT 31 du lundi au samedi vers 16h00. [Déplacements]

Dans cet extrait, le titre de section est constitué d'un groupe nominal qui est un terme complexe du domaine, repris immédiatement par un SN démonstratif *ce bulletin*, qui réalise là une réduction du terme complexe par effacement de l'expansion (cf. 3.1.2.2). Avant son apparition dans le titre de section, le terme complexe *bulletin prévisionnel* n'avait pas été mentionné dans le document.

De la même manière, une reprise peut intervenir à très grande distance d'une mention dans le titre, comme le montre (42), dans lequel le titre (la première ligne) comporte là encore un terme complexe, repris en fin de paragraphe sous la forme réduite *ce réseau*.

(42) UN RESEAU DE VOIES RAPIDES URBAINES SPECIFIQUE
TOULOUSE, capitale régionale de Midi-Pyrénées, au centre d'une agglomération particulièrement importante (plus de 650 000 habitants), se trouve à la convergence d'un réseau structurant étoilé permettant la desserte de territoires se situant bien au delà de la limite du Département de la HAUTE-GARONNE. Ce réseau étoilé aboutit sur les voies rapides urbaines de TOULOUSE constituées essentiellement : - des radiales situées à l'approche de l'agglomération - du périphérique proprement dit - de la liaison de l'Aéroport. - de la première section de la Rocade Arc-en-Ciel, doublant à l'ouest le périphérique. Ces voies rapides urbaines supportent un trafic particulièrement élevé (entre 60 000 et 130 000 véh / j), dont le maintien de la fluidité est primordial pour assurer la continuité du transit, une accessibilité correcte à l'agglomération Toulousaine, et un écoulement satisfaisant du trafic urbain. La gestion de ces voies étant multiple (ASF pour les autoroutes concédées, la DDE pour la voirie nationale et la voirie départementale), il est très vite apparu la nécessité d'instituer une coordination entre les exploitants, de façon à pouvoir optimiser l'usage de **ce réseau** stratégique : c'est ainsi qu'a émergé, dès la fin des années 80, l'idée du projet ERATO (Exploitation des Rocades de l'Agglomération TOulousaine). [Déplacements]

La récurrence de ces contextes qui voient une mention dans un titre de section pouvoir être ensuite réduite dans le texte sans autre forme d'introduction ou de consolidation du référent dans le discours s'est avérée intrigante. Ces configurations textuelles, apparemment favorables aux anaphores et à la réduction, interrogent sur les propriétés et les fonctions des titres de section : sont-ils des sites textuels particuliers, disposant de propriétés qui leur sont spécifiques et que ne présente pas le corps du texte ? Que sait-on des fonctions qu'ils remplissent dans les textes ? Comment appréhender et décrire ces fonctions ? Avec quelles « poignées » théoriques, à partir de quels traits ? Y a-t-il une littérature qui fournit un cadre d'analyse ?

4.1.1 Recherches sur les titres de sections (intertitres)

À l'orée des années 2000, peu de travaux linguistiques portent sur les titres de sections. Il y a des études de titres, mais sur les titres uniques d'une œuvre ou d'un texte : titres de tableau (Bosredon, 1997), titres de livres ou d'œuvres littéraires (Hoek, 1981), titres de presse (Sullet-Nylander, 2002). C'est là la relation de dénomination entre le titre et l'objet titré qui retient l'attention, loin donc des questions que j'ai évoquées, spécifiquement tournées vers l'insertion du titre dans le texte lui-même et vers sa contribution à la construction du texte comme texte.

Les recherches qui s'intéressent aux titres de section sont essentiellement le fait des psychologues. En particulier, R. et E. Lorch examinent l'impact des titres (*headings*[50]) dans des textes explicatifs sur des tâches de rappel ou de résumé (Lorch & Lorch, 1996). La question n'est pas de cerner la fonction de l'intertitre dans le texte, mais de mesurer ses effets du point de vue du traitement et de la compréhension du texte. Les recherches ne visent pas tant ce que dit ou fait l'intertitre que son éventuel rôle de facilitateur de la compréhension et de la mémorisation. L'enjeu peut en effet être fort : dans certains textes explicatifs ou procéduraux, la compréhension et le respect des consignes peuvent être une question de vie ou de mort, toute formulation ou mise en forme qui concourt à les maximiser doit donc être connue et exploitée (Cellier & Terrier, 2001 ; Heurley, 2001 ; Heurley & Ganier, 2006). Les psychologues travaillent ainsi le versant cognitif et non linguistique du titre de section. On verra plus loin que la rencontre avec le Modèle d'Architecture Textuelle, que je présente dans la section suivante (4.1.2, page 82), enrichit cette approche.

D'un autre côté, des travaux sur le discours s'intéressent précisément à la construction du texte et en particulier à ce qui concourt à lui donner organisation et structure. Charolles (1988) notamment identifie quatre plans d'organisation textuelle (période, chaine, portée, séquence) induisant chacun leur propre unité d'organisation. Mais ces plans jouent au niveau de phrases contigües et en régissent le regroupement ou le découpage, ils ne semblent donc pas aptes à fournir le cadre théorique nécessaire aux titres de sections. Poursuivant ses travaux, Charolles (1997) élabore un modèle d'encadrement du discours pour rendre compte de l'indexation de propositions successives dans un même ensemble intégrateur qu'il nomme « cadre de discours ». L'intérêt majeur de cette notion de cadre de discours, qui a rencontré un vif succès chez les linguistes préoccupés d'organisation textuelle, est d'offrir un modèle pour un phénomène qui dépasse la limite de la phrase et dont (43) donne un exemple.

(43) **Dans toutes ces cultures anciennes**, l'art n'a pas encore acquis sa totale autonomie par rapport au contexte culturel dont il dépend pour exister. Malgré le laminage de la modernité, il ne s'est pas laïcisé comme en Occident. Il reste englobé par le religieux, ce qui n'exclut pas une dimension esthétique propre, y compris endogène. Habitées à l'extrême, statues et peintures finissent par être vénérées par le dévot comme s'il s'agissait de Dieu lui-même. Elles constituent des points de passage entre le monde des humains et celui des dieux[51]. [Termith - anthropologie]

Le segment « dans toutes ces cultures anciennes » étend sa portée au-delà de la phrase qui l'accueille, jusqu'à la fin du paragraphe ici reproduit, et constitue un cadre qui intègre toutes les propositions exprimées, qui sont donc vraies « dans toutes ces cultures anciennes ».

50 La langue anglaise a cela de commode qu'elle permet de distinguer entre le titre général – *title* – et le titre de section – *heading* – et donc de pouvoir désigner précisément ce dont on parle. Pour éviter à la fois la répétition fastidieuse du terme complexe *titre de section* et une réduction source d'ambigüité, la dénomination *intertitre* a été progressivement préférée dans nos travaux, je l'adopte ici désormais.

51 Gérard Toffin (2009). Exposer/Voir. *L'Homme* 189, 139-164.

Ces notions de cadre et de portée ont l'avantage de fournir un modèle explicatif pour les phénomènes que j'ai évoqués : le fait qu'un terme complexe mentionné dans un titre puisse être repris par réduction à grande distance, le fait que ce que dit l'intertitre puisse servir d'interprétant pour ce qui est dit dans la section. Je montrerai plus loin qu'en effet, certains intertitres fonctionnent tout à fait comme des cadres de discours. En outre, ils ont en commun de jouer un rôle d'indexation, c'est-à-dire d'être tournés vers l'aval du texte (sur la distinction amont/aval, cf. Charolles & Péry-Woodley, 2005): ce qui suit le titre ou le cadre de discours. Mais malgré ces notions d'indexation et de portée, les cadres de discours sont envisagés dans une perception encore assez locale du phénomène linguistique alors que les titres de sections jouent un rôle au niveau du texte dans son entièreté.

C'est, là encore, la collaboration avec les chercheurs en informatique de l'IRIT qui s'est révélée fructueuse. Nous étions impliqués ensemble dans le projet Cognitique « Visualisation dynamique de textes »[52], piloté par C. Jacquemin, du LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, Paris-Sud). L'un des objectifs du projet était de proposer une visualisation par ordinateur qui facilite la navigation et l'appréhension du contenu de textes longs. Notre collègue de l'IRIT, Jacques Virbel, développait un modèle visant à rendre compte de la structure du texte, qu'il a nommé le Modèle d'Architecture Textuelle (Luc & Virbel, 2001). J'ai notamment utilisé ce modèle en collaboration avec M. Mojahid et L. Sarda (2001) pour un premier travail sur les structures textuelles. Dans la mesure où ce modèle répond à une partie des besoins descriptifs pour les titres de sections, une présentation se justifie. Elle sera toutefois succincte et simplifiée – peut-être même trop simplificatrice –, je renvoie donc à l'article cité ainsi que par exemple à (Virbel *et al.*, 2005) pour des compléments et une « mise en fonctionnement » du modèle.

4.1.2 Le Modèle d'Architecture Textuelle

Ce modèle est bâti sur deux fondements essentiels : i. un texte écrit présente des propriétés de mise en forme matérielle qui sont signifiantes et qui traduisent une intention de l'auteur à l'égard de son texte ; ii. la réalisation de l'intention de l'auteur appartient à la même classe d'équivalence qu'un énoncé métadiscursif qui expliciterait cette intention (l'arrière-plan théorique est ici l'approche mathématique du langage de Z. S. Harris[53]). Reprenons ces deux propositions.

L'apport essentiel de ce modèle est de prendre la dimension matérielle du texte au sérieux : loin de penser les propriétés typographiques et dispositionnelles du texte comme un habillage esthétique, un ornement qui viendrait agrémenter un message constitué en-dehors de cette matérialisation, la mise en forme matérielle du texte est pensée comme une des composantes de sa signification, au même titre que le sens des mots qui le constituent. Cette mise en forme matérielle agit par le contraste : c'est le fait par exemple d'utiliser une police de caractère différente, une taille différente, de l'italique, du gras, du soulignement, de mettre des numéros ou des tirets à certains endroits, de placer au centre de la ligne, en haut ou en bas de la page, qui va différencier des titres, des notes, des légendes, des exemples, des citations, des définitions, des axiomes, des énumérations, etc. Je viens ici de donner quelques exemples d'*objets textuels* constitués par cette mise en forme différentielle qui joue donc sur les propriétés *typo-dispositionnelles* de ces objets textuels pour les créer.

52 https://perso.limsi.fr/jacquemi/COGNITIQUE02/ page consultée le 28/07/2017.

53 Harris, Z.S. (1971). *Structures mathématiques du langage.* Paris, Dunod.

Pour manifester ces propriétés typo-dispositionnelles, J. Virbel exploite une technique d'« image de texte » qui donne à voir les éléments pertinents de la matérialité du texte. J'emprunte à (Luc & Virbel, 2001 : 103) l'image de texte qui va me permettre d'expliciter la seconde proposition. Cette image (Figure 6) veut représenter le début d'un chapitre d'ouvrage et se « lit » ainsi :

> Dans cette notation des structures textuelles par une « image de texte », le cadre représente un début de page, les espaces haut, gauche et droit, les marges, les chaînes interprétées ont leurs valeurs habituelles à l'échelle près, et Mxxxx note une chaîne quelconque débutant par une majuscule. (Luc & Virbel, 2001 : 105, note 1).

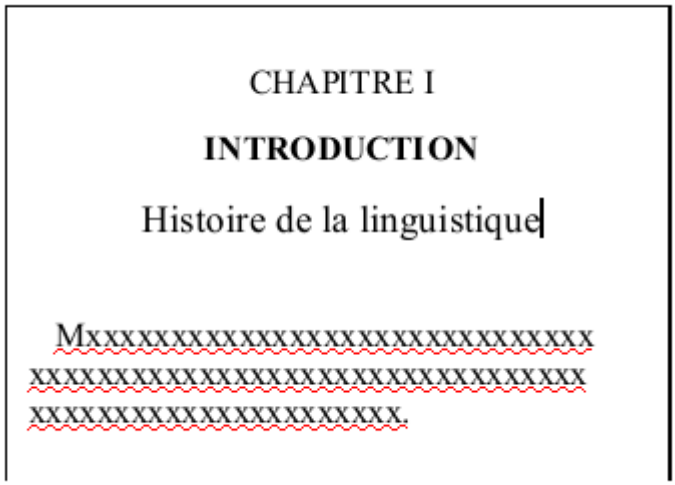


Figure 6 : *Image de texte*

Pour les auteurs, la mise en forme ici illustrée entretient une relation d'équivalence avec la phrase P1 : « *J'introduirai ce premier chapitre par une histoire de la linguistique* », phrase métatextuelle en ce qu'elle exprime l'action de l'auteur sur son texte. D'autres phrases métatextuelles (ou métaphrases dans la terminologie de Luc et Virbel) seraient « *je découpe le texte en chapitres* » « *j'attribue le numéro 1 à ce chapitre* », etc., « je » étant l'auteur du texte.

L'ensemble du texte peut ainsi être décrit par un petit nombre de métaphrases qui l'organisent, un *prototexte* dont la particularité est qu'il explicite toutes les actions de création du texte. La réalisation effective du texte final résulte d'une opération de *transformation / réduction* au sens de Harris, réduction qui laisse diverses traces pour les éléments réduits – traces matérialisées par l'italique, le gras, les indentations, etc. Un schéma proposé par (Lemarié *et al.*, 2008) éclaire parfaitement ce point, je le reproduis dans la figure 7. Il montre comment, à partir du même prototexte (haut du schéma), deux versions différentes de texte peuvent être dérivées.

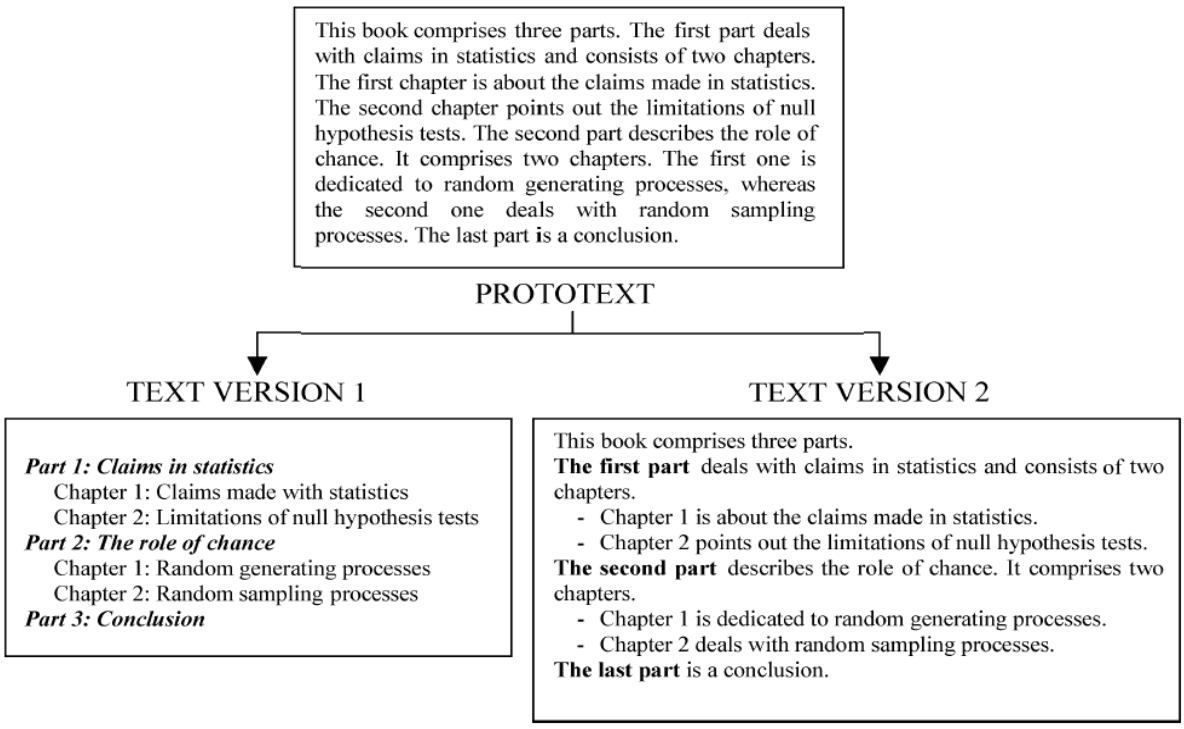


FIGURE 2. A prototext and some alternative realizations as actual texts.

Figure 7 : *Illustration de la notion de prototexte par (Lemarié et al., 2008 : 30)*

De la même manière, l'image de texte présentée dans la figure 6 est une transformation de la phrase métatextuelle P1. Il en ressort que les objets textuels créés par les contrastes de mise en forme matérielle peuvent donc être considérés comme équivalents à des formes discursives plus développées. Pour mieux l'illustrer, prenons pour exemples deux formes de définition :

(44) En géométrie euclidienne, un triangle est une figure plane, formée par trois points appelés *sommets*, par les trois segments qui les relient, appelés *côtés* [wikipédia[54]]

(45) **Triangle**
GÉOM. Polygone à trois côtés. [TLFi[55]]

Indépendamment du caractère laconique de la définition fournie par le TLFi, les deux formulations sont fonctionnellement équivalentes : elles indiquent le terme défini (triangle), l'univers de référence (la géométrie) et les éléments de définition. Toutefois, en (44), la mise en relation de ces éléments est assurée de façon linguistiquement explicite, par des moyens verbaux, alors qu'en (45), ce sont des moyens typo-dispositionnels qui remplissent le même rôle. La forme adoptée dans (45) peut être considérée comme une transformation de la forme discursive développée de (44) - qui serait elle-même une réduction d'une métaphrase dont la partie performative pourrait être « *j'inclus une définition du triangle* ». Cette équivalence est cruciale, et c'est pourquoi je la souligne, car c'est en vertu de cette équivalence que l'on est fondé à considérer que les caractéristiques typo-dispositionnelles des textes appartiennent à la sphère du linguistique et constituent donc à bon droit un objet d'étude légitime pour les linguistes.

54 https://fr.wikipedia.org/wiki/Triangle consulté le 28/07/2017

55 Trésor de la Langue Française Informatisé http://www.cnrtl.fr/definition/triangle consulté le 28/07/2017

Le MAT offre ainsi un modèle avec lequel saisir les fonctions des intertitres dans les textes, non seulement à l'égard du contenu titré, mais dans l'économie globale de l'architecture textuelle. Cependant, ainsi que le soulignent à juste titre Luc et Virbel (2001), le MAT ne permet pas de rendre compte de toutes les dimensions de signification des objets textuels, ici des intertitres, car les objets textuels sont aussi des objets discursifs. En d'autres termes, parallèlement à la matérialité du texte et à l'information que cette matérialité même délivre, un objet textuel, du fait d'employer des mots, est aussi porteur du sens que véhiculent ces mots. En gardant le même exemple de la figure 6, les intertitres « Introduction » et « Histoire de la linguistique » fournissent un découpage du texte et, dans le même temps, indiquent au lecteur la fonction (être une introduction) et la thématique (faire l'histoire de la linguistique) de la section ainsi découpée. Les deux opérations – découpage + indications – sont simultanées et relativement indissociables même si, pour l'analyse, on peut être amené à examiner tantôt l'une, tantôt l'autre, comme on le verra dans la section 4.2.

Comment rendre compte de ce versant qu'à la suite de Halliday (Matthiessen & Halliday, 2009), nous avons appelé *idéationnel* ? « Nous », car l'aventure a débuté en collaboration avec J. Rebeyrolle, qui avait fait dans sa thèse sur la définition des observations sur les reprises à la suite de titres (2000 : 148), et L.-M. Ho-Dac, qui s'intéressait aux cadres de discours (Ho-Dac, 2007). La section 4.2 fait le point sur l'élaboration collective des observables aptes à permettre une description satisfaisante des fonctions des intertitres avec un « nous » qui réfère à ce trio, sauf indication contraire.

4.2 Les travaux du « trio » : élaboration d'une méthode et premiers résultats

Étant donné la rareté de travaux linguistiques sur les titres, le premier travail à mener à bien pour aborder cet objet d'étude nouveau dans le champ est d'élaborer un cadrage apte à fournir les « poignées théoriques »[56] pour saisir cet objet et, avec ce cadrage, une méthodologie pour en produire une description.

Le MAT (Modèle d'Architecture Textuelle) donne une assise en permettant de considérer un intertitre comme la trace d'une opération de construction du texte. Cependant, même dans le modèle SARA (Lemarié *et al.*, 2008), proposé ensuite par une collaboration entre des psychologues et J. Virbel, l'auteur du modèle, les fonctions envisagées restent au niveau de la construction du texte et, quoique le modèle décrive ces fonctions textuelles d'une façon d'autant plus convaincante qu'il rejoint nos observations, il laisse sur sa faim en ce qui concerne le niveau du discours. S'agissant de la dimension idéationnelle, *i.e.* le contenu du texte, le modèle envisage l'intertitre comme annonçant le *topic* de la section, dans une relation telle que la section serait « à propos » (*about*) de ce dont « parle » l'intertitre (Lemarié *et al.*, 2008 : 33, Tableau 1)[57].

56 L'expression est de Marie-Paule Péry-Woodley, qui conduisait pour l'ERSS les travaux du projet « Visualisation dynamique de textes ». Les débuts de ces recherches sur les titres de section doivent beaucoup à ses remarques éclairées et éclairantes.

57 Il faut souligner toutefois de la part des auteurs une évolution vers une prise de conscience de la simplification qu'ils ont inconsciemment opérée quand ils construisaient des recherches sur les titres et leurs effets : « psychological researchers appear to have taken a simplistic view of titles and headings, ignoring their capacity to fulfil and possibly combine several functions. The empirical literature should therefore be critically revisited, as it may overgeneralize findings which apply only to certain types of headings. » (Lemarié, Lorch & Péry-Woodley, 2012, § 8)

La notion de *topic* et d'*à-propos* est de celles qui sont susceptibles de donner lieu à des perceptions variées et à des jugements divergents (cf. 2.2.2). Qu'est-ce que signifie concrètement « être à propos », comment cela se traduit-il, et surtout comment cela s'évalue-t-il ? En effet, si l'on reprend l'exemple (42), page 80, montrant l'intertitre « UN RESEAU DE VOIES RAPIDES URBAINES SPECIFIQUE », peut-on dire que la section est *à propos* d'un réseau de voies rapides urbaines spécifique ? Et si l'on reprend la figure 5, peut-on dire que la section 2 du texte serait *à propos* « Des corpus représentatifs : de quoi ? » ou de la question posée par ce titre ? Ces deux exemples suffisent déjà à montrer le flou inhérent non pas à la notion, assez circonscrite dans (Lambrecht, 1994), mais à son opérationnalisation pour une description. Nous avons donc abandonné l'idée d'utiliser cette notion pour la caractérisation des fonctions des intertitres et nous nous sommes efforcées de bâtir le système autant que possible cohérent et objectivant de description que je présente maintenant.

4.2.1 Une description à base de traits formels et d'annotation

L'un de nos objectifs en abordant la problématique des intertitres concernait le traitement automatique de la langue naturelle[58] (TAL). Le projet que j'ai déjà évoqué, « Visualisation dynamique de textes », supposait en effet une analyse automatique des structures textuelles pour en proposer une visualisation adaptée. La description à produire était ainsi partiellement guidée par cet objectif applicatif. Notamment, l'horizon de traitements automatiques a impliqué de privilégier des traits calculables par un ordinateur, c'est-à-dire des traits formels.

Mais le choix de concevoir les outils descriptifs à base de traits formels n'était pas dû uniquement aux exigences de TAL, il correspondait surtout au souci d'éliminer autant que possible la subjectivité et l'interprétation dans l'appréhension du rôle de chaque titre. Nous étions pleinement conscientes des écueils liés aux jugements individuels des linguistes sur les données et ambitionnions de les éviter en recourant à des observables peu sujets aux divergences d'appréciation. J'ai regretté par la suite que cette conscience n'aille pas jusqu'à mettre en place une étape formelle d'évaluation de nos jugements sur les données – comme elle a été menée pour l'annotation en texte des procédures, cf. section 2.2.2, page 44 –, je ne suis donc pas en mesure d'indiquer une mesure de notre accord.

La démarche, exposée dans (Jacques, Rebeyrolle & Ho-Dac, 2004) [9], a donc été la suivante :

- formulation d'hypothèses théoriques générales sur les fonctions des intertitres ;

- choix d'observables cohérents avec ces hypothèses ;

- annotation / codage de chaque titre du corpus pour ces observables ;

- analyse quantitative fondée sur des mesures statistiques des corrélations de traits – à la Biber (1988) ;

- interprétation qualitative des regroupements révélés par l'analyse statistique.

Nous avons constitué un corpus de 1041 intertitres en joignant une partie des corpus sur lesquels chacune avait mené sa thèse de doctorat : textes professionnels dans le domaine des Déplacements (le corpus évoqué au chapitre précédent), articles scientifiques de Géopolitique[59], articles scientifiques du domaine de l'Ingénierie des Connaissances.

58 Je rappelle que je consacrerai le chapitre 5 plus spécifiquement aux recherches qui concernent le traitement automatique des langues.

59 Ce corpus a été constitué par Mai Ho-Dac, qui l'a « aspiré » sur le site web de l'IFRI (Institut Français des Relations Internationales) et a obtenu de cet institut les autorisations pour l'exploiter et le rendre disponible. Il a reçu nombre d'enrichissements, consultables et téléchargeables sur la page Redac du

Deux hypothèses principales ont constitué la source de la réflexion :

1) La forme et la fonction des intertitres varient selon le genre de discours ;

2) Les intertitres ne sont pas seulement des « balises » dans les textes écrits, ils ont une fonction de segmentation, de hiérarchisation du texte et **en plus** ils contribuent à l'organisation du contenu du discours.

L'hypothèse 1 devait se tester par la confrontation des trois sous-corpus : si nous avions raison, nous devions trouver plus de ressemblance entre les intertitres des articles scientifiques qu'entre ceux-ci et les intertitres des textes professionnels. J'y reviens dans la section 4.2.2.4, p. 96.

Pour l'hypothèse 2, il fallait aller plus loin et envisager la façon dont les intertitres contribueraient au contenu du discours. Nous avons donc affiné cette hypothèse en supposant aux intertitres un rôle cohésif et en en déduisant les observables à prendre en considération. Si l'on suit (Charolles, 1978), une part de la cohésion textuelle est assurée par la récurrence d'éléments (méta-règle de répétition, page 14). Il semblait donc pertinent de s'intéresser à la répétition des unités lexicales formant l'intertitre et de caractériser finement cette répétition. Pour la facilité de l'annotation, celle-ci s'est effectuée à travers un formulaire[60] dont la figure 8 montre une copie d'écran. Le fait de disposer d'une interface visuelle est une aide cognitive non négligeable qui permet au linguiste d'allouer ses ressources mentales aux données et non à l'outil. C'est la voie suivie par d'autres projets de recherche qui s'intéressent aussi aux phénomènes textuels, par ex. (Landragin, 2016). Ce formulaire a donc été conçu pour afficher, en même temps que le titre, une part de son contexte et pour donner, si nécessaire, accès à l'ensemble du texte (bouton rouge *Voir contexte ou reprise*).

La partie gauche du formulaire, repérée par l'intitulé « Reprises » montre le système de traits tels que nous l'avons fait fonctionner. Il résulte de phases d'annotation / enrichissement du modèle : au fur et à mesure que nous codions les intertitres, nous revenions sur nos choix de modèle pour les valider ou les affiner.

Le modèle s'est donc stabilisé avec 8 traits à apprécier, chacun ayant un nombre variable de valeurs (pour une description détaillée, cf. Jacques & Rebeyrolle, 2006) [18] :

1. le fait qu'il y ait ou non une reprise du titre (si la réponse était 'non', tous les traits suivants prenaient la valeur 'non concerné') ;
2. la forme de cette reprise ;
3. le fait que cette reprise soit ou non dispersée dans le texte de la section (par exemple, dans la figure 8, les unités lexicales du titre sont répétées à deux endroits différents : *approche* puis *texte*) ;
4. le lieu du texte où se produit la reprise ;
5. la position sujet de la reprise ;
6. la présence d'un autre titre suivant immédiatement l'intertitre à coder ;
7. le fait qu'il y ait une reprise dans cet autre titre ;

laboratoire CLLE-ERSS : http://redac.univ-tlse2.fr/corpus/geopo.html (page consultée le 4/08/2017).

60 Mai Ho-Dac et moi-même avons mis en commun nos connaissances des bases de données et du langage Visual Basic pour élaborer et la base de données sous-jacente et les différents formulaires.

8. le fait que la reprise ne répète pas exactement une unité lexicale du titre mais une conversion (par exemple *Repérage* / *repérer*).

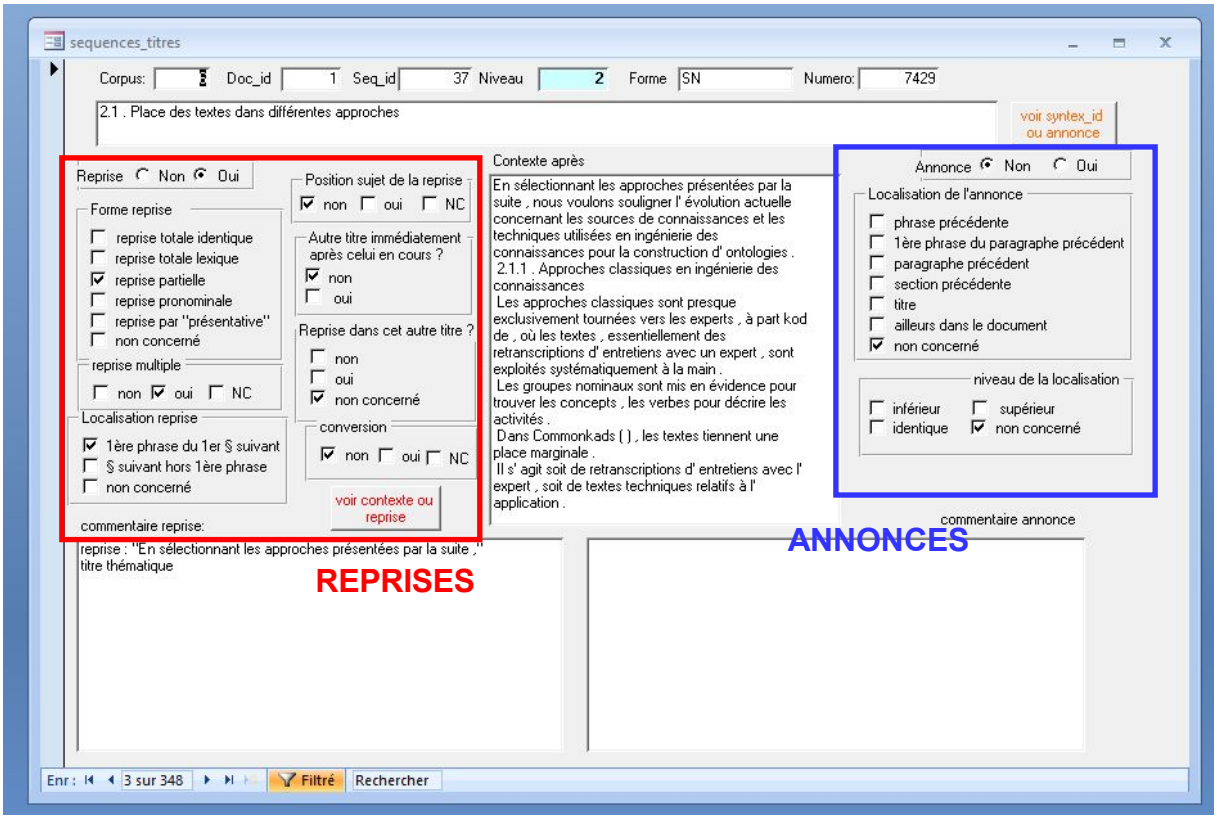


Figure 8 : *Formulaire d'annotation des intertitres sous l'angle de la cohésion*

La partie droite du formulaire est dédiée à une éventuelle annonce dans « l'avant-titre » : toute la partie du texte qui précède l'intertitre, sans limite d'empan. Nous prenions en considération ce qui précède l'intertitre, de la phrase placée immédiatement avant lui jusqu'au tout début du texte lui-même, pour y traquer toute mention qui anticipe l'intertitre, par exemple (46), où la phrase qui précède immédiatement l'intertitre 2.1.1 annonce cet intertitre et le suivant, 2.1.2.

(46) Les connaissances du domaine sont exprimées d'une manière déclarative au travers d'une **composante terminologique** et d'une **composante déductive**.
2.1.1 La **composante terminologique**
La composante terminologique de Carin-ALN comprend des définitions et des inclusions de concepts. […]
2.1.2 La **composante déductive**[61] [article scientifique – ingénierie des connaissances]

Je laisserai de côté pour le moment cette annonce, car nos analyses en ont fait peu de cas, j'y reviens dans des travaux et un projet plus récents (section 4.4).

Par ailleurs, un autre formulaire (figure 9) permettait de noter finement les caractéristiques de la forme des intertitres.

61 Reynaud C., Safar B., Gagliardi H. (2001) La représentation de l'ontologie du domaine dans le médiateur PICSEL, *12èmes Journées Francophones d'Ingénierie des Connaissances, IC'01,* Grenoble, 25-27 Juin 2001.

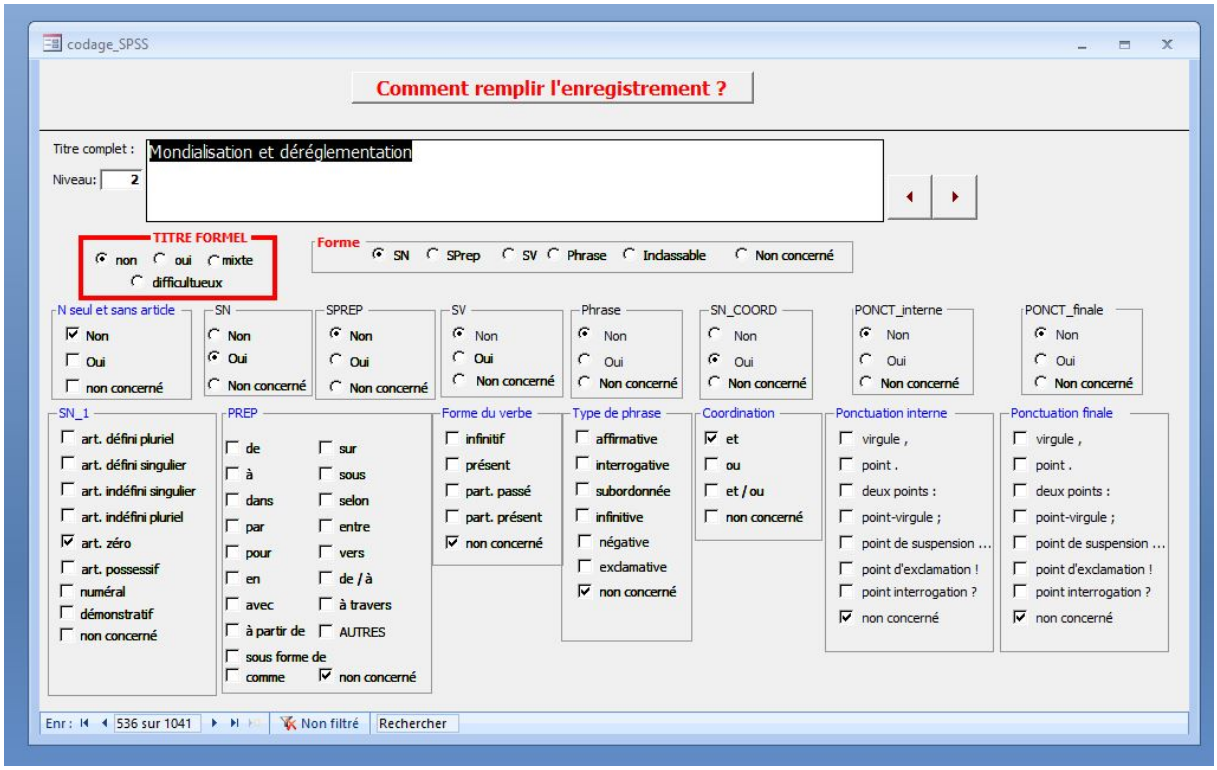


Figure 9 : *Codage des caractéristiques formelles des intertitres*

Nous avons notamment distingué ce que nous avons appelé des *titres formels* (titres fonctionnels, *functional headings* dans le MAT de Lemarié *et al.*, 2008, page 35) et les autres intertitres. Les titres formels sont les intertitres qui indiquent ce **qu'est** la section ou l'objet textuel qui suit, tels que *Chapitre*, *Introduction*, *Conclusion*, *Résumé*. La métaphrase sous-jacente à ces intertitres serait du type « l'unité suivante **est un chapitre** (ou une introduction ou une conclusion) » alors que la métaphrase sous-jacente à un titre tel que celui de la figure 8 « Place des textes dans différentes approches » serait « l'unité suivante **est à propos de** la place des textes dans différentes approches » (quoique *être à propos* soit une façon trop faible et trop vague d'indiquer la relation entre l'intertitre et l'unité qu'il intitule).

Certains intertitres peuvent être hybrides (selon la terminologie de Lemarié *et al.*, 2008), c'est-à-dire qu'ils peuvent présenter à la fois une partie fonctionnelle et une partie idéationnelle, comme par exemple : « Chapitre 2 - Ancrage théorique, données et méthodes ».

Le croisement de tous ces traits, formels et « cohésifs », a permis de mettre en évidence des rôles discursifs diversifiés pour les intertitres.

4.2.2 Résultats : des configurations de traits pour cerner des fonctionnements

Les résultats présentés ici combinent les résultats exposés dans (Ho-Dac, Jacques & Rebeyrolle, 2004 ; Jacques, 2005c ; Jacques & Rebeyrolle, 2006)[62] [8, 13, 18].

62 Un autre article (Rebeyrolle, J. et Jacques, M.-P. (2006). Étude en corpus de la fonction des intertitres dans la construction du discours. In: *3e Rencontre fribourgeoise de la linguistique sur corpus appliquée aux langues romanes*, *Université Albert-Ludwig de Fribourg en Brisgau* (Allemagne), 14-17 septembre 2006) aurait dû compléter ces travaux. Malheureusement, la publication des actes de la conférence a été abandonnée.

4.2.2.1 Segmenter donc rassembler et hiérarchiser

La première fonction de l'intertitre, indépendamment de son contenu ni même du fait qu'il s'agisse d'un intertitre fonctionnel ou d'un intertitre « idéationnel », est d'assurer une segmentation du texte. Il introduit des divisions de façon visible et découpe des sous-ensembles. Comme l'a judicieusement fait remarquer Péry-Woodley (2001 : 1), segmenter c'est à la fois diviser mais aussi rassembler. Une image de texte, reprise de (Ho-Dac *et al.*, 2004) [8], mettra cette fonction en évidence (figure 10). On y voit, littéralement parlant, que les intertitres matérialisent des délimitations et, ce faisant, matérialisent le regroupement des paragraphes et des subdivisions sous-ordonnées.

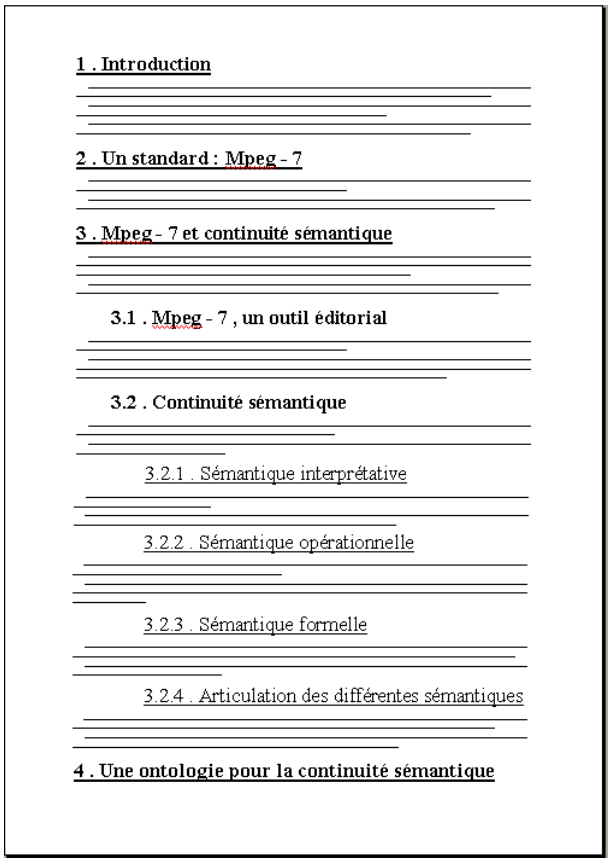


Figure 10 : *Image de texte des premières sections d'un article scientifique en Ingénierie des Connaissances*

Ce jeu de divisions / regroupements induit une hiérarchie : des sections sont enchâssantes, d'autres sont enchâssées. C'est ainsi la structure sémantico-logique du texte qui transparait à travers le système d'intertitres et de leurs relations d'un texte – ce que nous avons nommé la *titraille*.

Dans (Jacques, 2005c) [13], j'ai décrit plusieurs modes de relations possibles pour les unités qui sont « sœurs » au sein d'une même section englobante, et dont l'image de texte dans la figure 10 offre un aperçu. Sous le chapeau « 3.2. Continuité sémantique », les sections « 3.2.1 », « 3.2.2 », « 3.2.3 » déclinent chacune un type de sémantique. Elles sont donc en relation de parallélisme : les intertitres indiquent que chaque section est au même niveau logique que la précédente, les sections sont relativement interchangeables.

La section « 3.2.4 » en revanche se propose d'articuler les différentes sémantiques qui viennent d'être exposées. Elle entretient avec les précédentes une relation analogue à celle qu'occupe une conclusion à la fin d'un article ou d'un ouvrage : elle reprend des éléments développés

pour en offrir une nouvelle mise en perspective ou un prolongement. Elle ne pourrait donc se situer ailleurs qu'à la fin de la section enchâssante.

Dans le même ordre d'idées, des procédés anaphoriques dans les intertitres manifestent l'ordonnancement du raisonnement ou de l'exposition dans le texte. Prenons pour exemple la séquence d'intertitres en (47), la numérotation précise la hiérarchie, je supprime délibérément les sections pour focaliser sur les intertitres.

(47) 1.2. Tours "intransitifs"
1.2.1. Ça joue.
1.2.2. Autres exemples de forme intransitive[63] [Termith - linguistique]

L'adjectif *autres* de l'intertitre 1.2.2. positionne cette section comme venant nécessairement après une autre, qu'elle complète, précise, nuance…

Sans exprimer une relation aussi claire, les intertitres peuvent manifester une forme d'« entonnoir » pour le déroulement du texte, structure particulièrement présente dans les articles scientifiques des SHS (Sciences Humaines et Sociales) : chaque section prolonge un aspect évoqué dans la section précédente. Je reprends ici un exemple de (Jacques, 2005c) [13] :

(48) L'univers des think tanks, "seconde économie" de l'influence politique
La corruption des think tanks
Quelques exemples de la corruption des think tanks [article scientifique - géopolitique]

et le schéma qui rend compte de cet emboitement (figure 11) :

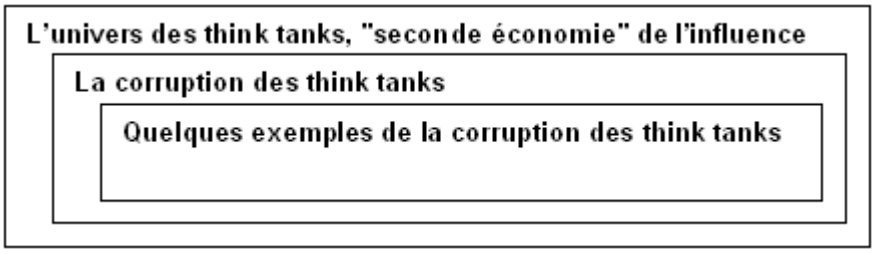


Figure 11 : *Relation d'emboitement entre sections de même niveau*

Ces différents extraits témoignent de la façon dont les intertitres, considérés en-dehors même de leur intégration dans le texte, donc des fonctions cohésives que je vais exposer par la suite, construisent l'armature logique du texte, en le divisant et en répartissant les divisions ainsi créées en blocs, invitant le lecteur à traiter le contenu selon ces blocs.

Un autre aspect notable, souligné aussi par (Lemarié *et al.*, 2008), et qui transparait de ce qui précède est la capacité de l'intertitre à fournir un label pour ces divisions du texte. Sont ainsi facilitées les opérations de renvoi ou d'anticipation telles qu'en (49) et en (50), ci-après, qui concourent au « tissage » du texte, en particulier dans les articles de recherche scientifique – ces deux exemples sont d'ailleurs extraits du corpus Termith.

(49) Comme nous l'avons dit plus haut **(2.1)**, [Termith - linguistique]

(50) comme nous le détaillerons au **§ II. 4** [Termith - linguistique]

Découper, délimiter, introduire des regroupements, des parallèles, des emboitements sont donc des opérations réalisées par les intertitres, indépendamment des modalités de leur insertion dans

63 Romero-Lopes Marcia Cristina (2002). Identité et variation du verbe jouer. *Langue française*, 133, pp. 63-73.

le texte. La face idéationnelle de l'intertitre (ce que les mots qui le composent signifient) établit des liens entre les sections et crée une macrostructure logique. Changeons maintenant de perspective et entrons dans l'univers du texte pour tenter de saisir la façon dont les intertitres contribuent à construire cet univers.

4.2.2.2 Rôle des intertitres dans la construction du sens du texte

Le modèle d'annotation élaboré correspond, je le rappelle, à l'ambition de cerner le rôle cohésif de l'intertitre. Dans nos premiers travaux (Ho-Dac *et al.*, 2004) [8], nous avons distingué trois modes de participation de l'intertitre à la construction du « monde du texte » (Werth, 1999) et à la cohésion textuelle, ce que nous avons appelé son *implication* :

- implication « zéro »,

- implication référentielle,

- implication thématique.

L'implication zéro est celle des intertitres que nous nommions formels – titres fonctionnels pour (Lemarié *et al.*, 2008), c'est-à-dire les intertitres tels que *Chapitre* ou *Partie*. Leur fonction est strictement textuelle dans le sens où ils assurent le découpage du texte et indiquent la nature de la partie découpée mais ne disent rien de plus de son contenu.

Les deux autres modes d'implication construisent un continuum sur lequel les différents intertitres idéationnels se répartissent. Chaque mode d'implication est corrélé à des configurations des traits que nous avons codés (cf. figure 8, plus haut), qui constituent alors des indices du mode d'implication. Je précise maintenant pour chaque pôle ses caractéristiques, des exemples et les traits pertinents qui le marquent.

1. Intertitres à implication référentielle

Ce sont les intertitres qui introduisent ou réactivent un référent du discours. Ils sont de ce fait l'un des maillons d'une chaine de référence (Schnedecker & Landragin, 2014), parfois même le premier maillon. (51), (52) et (53) en donnent des exemples – pour la lisibilité des exemples, je mets en caractères gras tous les éléments des reprises :

(51) 5.3 **La réutilisation**
L'une des techniques proposées pour faciliter le processus de modélisation, en ingénierie des besoins comme en ingénierie des connaissances, est **la réutilisation** de modèles. **Elle** devient un objectif prépondérant[64]. [article scientifique – ingénierie des connaissances]

(52) 10.1.2 **LES EQUIPEMENTS COMMUNS**
Les équipements communs sont, a contrario, tous les équipements généralement situés hors des réseaux routiers des exploitants et **dont** l'utilisation n'est justifiée que par la mise en oeuvre des actions ERATO. On **y** retrouvera donc essentiellement le réseau et les matériels de transmission reliant les PC propres de chaque exploitant au Centre de Contrôle de Trafic (CCT) ERATO ainsi que ce dernier et ses équipements. [Déplacements]

(53) **LES VARIABLES DÉPENDANTES**
Elles consistaient en un questionnaire que les participants recevaient après avoir visionné l'interview[65]. [Termith - psychologie]

64 Tort F., Teulier R., Grosz G., Charlet J. (2000). Ingénierie des besoins, ingénierie des connaissances : similarités et complémentarités des approches de modélisation. *IC'2000*, Toulouse, pp. 263-275.

En (51) et en (52), l'intertitre constitué d'un syntagme nominal est intégralement repris dans la première phrase de la section, les expressions en caractères gras montrent les autres maillons de la chaine qui concerne ce référent. En (53), la reprise, immédiate aussi, prend la forme d'un pronom personnel, mais la chaine de références s'arrête là.

La configuration de traits qui caractérise ces intertitres combine l'immédiateté de la reprise – celle-ci intervient dès la première phrase de la section – et deux formes de reprise privilégiées : soit une répétition identique du titre, soit une reprise pronominale, comme en (53). La forme même de l'intertitre entre aussi en considération puisque sont référentiels principalement les intertitres de type SN, formés d'un bloc unique, c'est-à-dire à l'exclusion des intertitres qui contiennent une coordination ou une ponctuation introduisant une segmentation dans l'intertitre – un intertitre tel que « *MPEG-7 et continuité sémantique* » de la figure 10 est ainsi formé de deux blocs.

Sur le plan cognitif, si l'on se place du point de vue du lecteur, l'instruction véhiculée par ces intertitres référentiels correspond à la mise au premier plan du référent mentionné dans le titre, afin de le rendre accessible pour la suite du texte.

2. Intertitres à implication thématique

À l'autre extrémité du continuum, les intertitres thématiques indiquent ce dont on va parler. Ce faisant, ils inscrivent le texte titré dans un domaine d'activité, un domaine de connaissance, un point de vue, une situation spatio-temporelle, etc., spécifiques. Certains se rapprochent ainsi des cadres de discours (Charolles, 1997) que j'évoquais en début de chapitre.

(54) A- Après le rapport **Rumsfeld**, les **réorganisations** en cours
Les affaires spatiales aux Etats-Unis sont affectées par une **réorganisation** institutionnelle, qui trouve son origine dans les propositions d'une récente commission parlementaire. L'actuel secrétaire à la Défense Donald **Rumsfeld** a présidé […] une commission indépendante…[66] [article scientifique – géopolitique]

(55) 4.2 **Tracer l'activité** : **les observables**
Une des difficultés pour construire de tels modèles est la diversité **des observables**, **traces** de **l'activité**. C'est le cas par exemple des calendriers de pâturage étudiés par […], qui inscrivent la **trace** de **l'activité** dans le temps et l'espace d'un éleveur[67]. [article scientifique – ingénierie des connaissances]

Les reprises après ces intertitres peuvent conduire à parsemer le texte des unités lexicales présentes dans l'intertitre, ce qui a pour effet une omission possible d'une partie de l'intertitre – en (54), seuls deux mots sont repris tels quels – ou même la conversion grammaticale d'un des mots de l'intertitre – en (55), *tracer* est repris par *traces* puis *trace*. Les intertitres thématiques ne participent pas aux chaines de référence du texte. Ils sont les parangons des titres auxquels Lemarié *et al.* (2008) font référence en indiquant que le contenu titré est *à propos* de ce qu'énonce le titre : en effet, dans ces deux cas, la section est à propos des réorganisations (54) ou à propos des observables qui constituent les traces de l'activité (55).

---

65 Cicotti S. (2004). Les orateurs ont-ils intérêt à s'épiler les sourcils ? Influence d'indices non verbaux sur les processus persuasifs. *L'année psychologique*, 104, n°2. pp. 227-247.

66 Nardon L. (2002). Militarisation, gestion et coopération : l'administration Bush et l'espace. IFRI https://www.ifri.org/sites/default/files/atoms/files/nardonbush0402.pdf (consulté le 3/08/2017).

67 Teulier R., Girard N. (2005). Modéliser les connaissances pour l'action dans les organisations. In Teulier R., Charlet J., Tchounikine P. (éds.) *Ingénierie des Connaissances*, L'Harmattan, pp. 389-412.

Les traits typiques de ces intertitres résident précisément dans cette dispersion, souvent tout au long d'une section, des répétitions des unités du titre et dans l'absence de chaine de référence à partir des référents construits dans le titre. Nombreux sont les intertitres thématiques qui, comme dans ces exemples, présentent une segmentation interne, soit par une ponctuation (virgule, point-virgule, deux-points, tirets, etc.), soit par un coordonnant tel que dans « *4.2. Adhésion et observance* ».

Sur le plan cognitif, contrairement aux intertitres référentiels qui visent à attirer l'attention du lecteur sur un ou des référents du discours particulier(s), les intertitres thématiques l'invitent à mobiliser certaines de ses connaissances d'arrière-plan en vue d'intégrer le propos à suivre.

3. Entre les deux pôles

Nous avons décrit là les intertitres se situant clairement à chacun des pôles mais, je le rappelle, il s'agit avant tout de continuum. La plupart des intertitres, en particulier dans les textes scientifiques, jouent un rôle à la fois sur la gestion des référents et sur la thématique du discours, ce qu'illustre bien (56).

(56) La **singularité** de **la contestation à la française**

> **La contestation française** de la mondialisation n'est pas tant **singulière** par ses caractéristiques, que par son influence notable sur la société et sur le débat politique. Cette **contestation à la française** est née, en grande partie, dans le sillage du débat sur Maastricht, mais aussi de la réapparition de mouvements sociaux importants, avec les actions des "sans" (logement, travail, papier) et surtout les grèves du secteur public fin 1995. La création, en juin 1998, d'ATTAC et son succès rapide, l'intense campagne menée par la Coordination contre l'AMI (Accord Multilatéral sur l'Investissement négocié à l'OCDE) et le démontage du restaurant McDonald's à Millau par des militants de la Confédération paysanne en août 1999 en réaction à la décision américaine de surtaxer des produits agricoles français, font de la France à travers ses figures médiatiques, notamment celle de José Bové, l'un des hauts lieux de la lutte contre la "mondialisation libérale"…[68] [article scientifique – géopolitique]

Une partie de l'intertitre constitue un maillon de chaine référentielle, mais la tête du SN que forme l'intertitre est reprise sous forme adjectivale comme attribut de ce référent. L'intertitre « joue sur les deux tableaux » : braquer le projecteur sur un référent et préciser le point de vue par lequel ce référent va être traité dans la section.

Nous avons dans (Ho-Dac *et al.*, 2004) [8] envisagé cinq gradations sur ce continuum entre implication référentielle et implication thématique, mais ce découpage est relativement arbitraire car la perception de la fonction prédominante reste floue et subjective. C'est pourquoi nous avons poursuivi dans (Jacques & Rebeyrolle, 2006) [18] le travail sur la validation statistique de notre modèle.

4.2.2.3 Validation statistique à partir d'analyse multifactorielle[69]

Notre objectif était d'assoir le modèle de fonctions proposé sur une mesure statistique. Celle-ci devait mettre en évidence ou balayer les regroupements opérés et les fonctionnements déduits à partir du codage de traits. Le recours à des mesures statistiques est une suite logique de l'ensemble de la démarche : se baser sur corpus, limiter la subjectivité et l'interprétation par un

68 Fougier E. (2003). La contestation de la mondialisation : une nouvelle exception française ? *Policy brief* n°2. IFRI.

69 Cette analyse a été pilotée surtout par J. Rebeyrolle, qui a manipulé le logiciel SPSS avec *maestria*.

choix de traits formels, mener des observations sur un grand nombre de données. La quantité d'intertitres considérés (1041) et de valeurs possibles pour les traits retenus exclut la perception de corrélations sans un outillage adapté. La mesure statistique apparait alors comme l'un des outils du linguiste de corpus, avec le concordancier et la base de données[70].

La mesure la plus adaptée aux données et à notre démarche de codage de traits est une analyse multifactorielle.

> Il s'agit en effet d'une analyse qui permet de tenir compte non du rôle des variables indépendamment les unes des autres mais de leur influence conjointe. Ce type d'analyse statistique permet de confirmer les oppositions posées théoriquement comme pertinentes en validant statistiquement ou non la pertinence des traits linguistiques considérés comme déterminants pour classer les titres et d'interpréter ces classements en termes de fonctions discursives. (Jacques & Rebeyrolle, 2006 : 7) [18]

Pour alléger le propos, je ne reproduis pas ici les différents tableaux de mesures, on pourra retrouver les détails de l'analyse statistique dans (Jacques & Rebeyrolle, 2006 : 8-11) [18], je reprends la figure la plus éloquente de ces calculs statistiques (figure 12), qui en présente la conclusion.

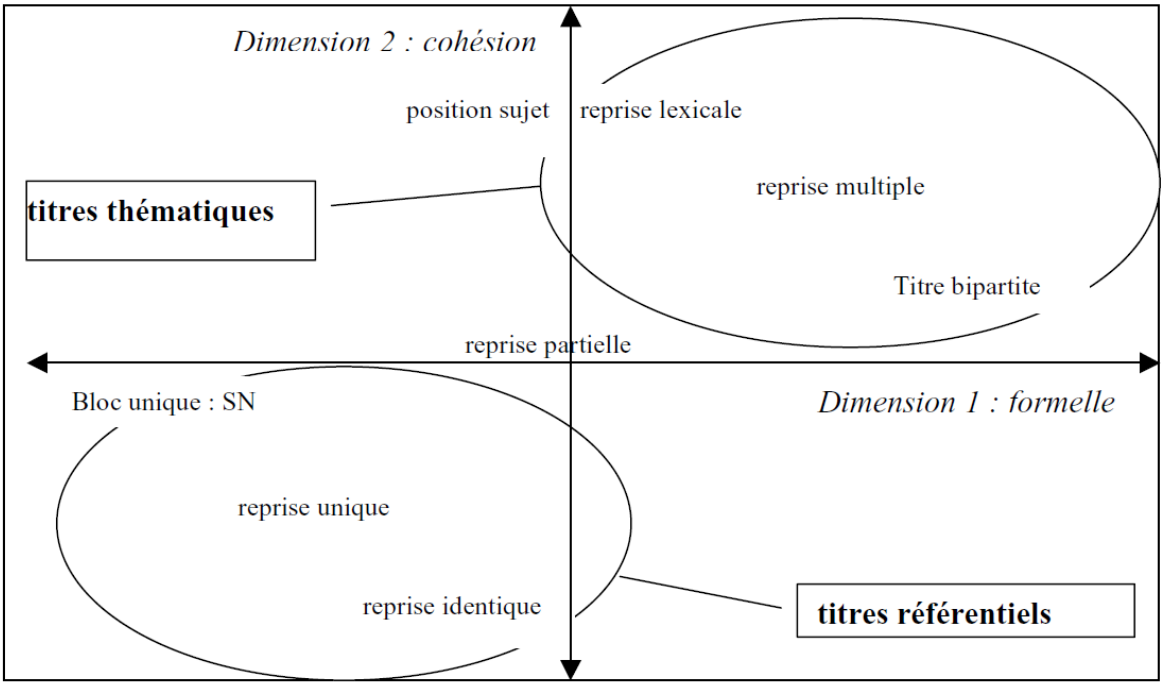


Figure 12 : *Opposition entre implication référentielle et implication thématique*

La figure 12 montre la projection sur un plan des deux dimensions rendant compte de la variance de nos données, dimensions que nous avons interprétées comme traitant de l'aspect formel et comme traitant de l'aspect cohésif. Elle montre la position des variables qui contribuent le plus fortement à expliquer la variation. Nous traçons ensuite les deux cercles autour des regroupements que l'analyse statistique opère entre les sous-ensembles d'indices que nous avons considérés comme marquant l'un ou l'autre type d'implication.

70 C'est ce qui a d'ailleurs motivé Habert (2009) à rédiger un ouvrage pour introduire les bases de données auprès d'étudiants et de chercheurs de lettres et sciences humaines.

Hormis la position sujet, que nous pensions typique de l'implication référentielle et qui ne se trouve pas placée dans le même espace que les autres traits de cette implication, l'opposition est validée par l'analyse : on a bien d'un côté des intertitres plutôt bipartites, donnant lieu à une reprise éparpillée dans le texte de l'ensemble des unités lexicales de l'intertitre, et d'un autre côté des intertitres plutôt monobloc, constitués d'un SN sans marque de partition interne, et donnant lieu à une reprise à l'identique.

La position centrale de la reprise partielle est tout à fait cohérente avec ces regroupements. Une reprise partielle peut être soit la reprise de la tête du SN, avec une opération de réduction par effacement de l'expansion telle que je l'ai décrite au chapitre précédent (par ex. *les chantiers courants → ces chantiers*), assez typique de l'implication référentielle, soit la reprise d'une autre partie de l'intertitre, comme dans (54) qui voyait la répétition de « réorganisation » et de « Rumsfeld » tout en laissant de côté les autres unités lexicales de l'intertitre.

Cette analyse nous inviterait à affiner le modèle, d'une part, pour éclaircir la question de la fonction sujet de la reprise, d'autre part, pour mieux spécifier la nature d'une reprise partielle. En l'état, elle fournit tout de même une assise pour le modèle de fonctions que nous avons dessiné. Et elle donne aussi le moyen d'apprécier la variation des types d'intertitres selon le genre textuel.

4.2.2.4 Variations selon le genre

Je rappelle que les analyses présentées ont été menées principalement sur trois sous-corpus :

- certains des textes professionnels du corpus Déplacements présenté au chapitre 3,

- des articles scientifiques de géopolitique (voir note 59, p. 86),

- des articles scientifiques d'ingénierie des connaissances (thèse de J. Rebeyrolle, 2000).

On aura noté au passage des exemples provenant du corpus Termith, présenté lui aussi au chapitre 3, constitué d'articles scientifiques des SHS. Le fait de trouver des exemples dans des textes extérieurs aux corpus qui ont servi de supports à l'élaboration du modèle est un indice en soi d'une certaine validité du modèle – ce qui n'est pas négligeable pour une linguistique de corpus qui peut se voir opposer la contingence de ses observations, cf. chapitre 2. C'est aussi un indice du fait que les fonctions des intertitres sont corrélées au genre du texte, ce que confirme une procédure statistique complémentaire.

Dans (Ho-Dac *et al.*, 2004) [8], nous avions mis en évidence une corrélation entre genre et forme : même si la forme privilégiée des intertitres est dans les trois corpus le syntagme nominal, seuls les articles scientifiques réservent plus de 10 % d'occurrences aux autres formes, syntagmes prépositionnels (57), syntagmes verbaux (58), voire phrases (59), contre moins de 1 % pour le corpus Déplacements (2/348 intertitres).

(57) A- Après le rapport Rumsfeld, les réorganisations en cours[71] [article scientifique - géopolitique]

(58) 5.2.3 Assister la prise de décision collective[72] [article scientifique - ingénierie des connaissances]

71 Nardon L. (2002). Militarisation, gestion et coopération : l'administration Bush et l'espace. IFRI https://www.ifri.org/sites/default/files/atoms/files/nardonbush0402.pdf (consulté le 3/08/2017).

72 Darses F. (2001). Concevoir des systèmes à bases de connaissances destinés aux tâches de conception : préconisations ergonomiques. Conférence *Ingénierie des Connaissances*, Grenoble. NB : Je ne suis pas sûre à 100 % de cette référence.

(59) 3. Le système économique soviétique reste très présent dans le bassin Caspien[73] [article scientifique - géopolitique]

Or nous avons vu que la forme intervient comme facteur dans l'analyse statistique, qu'en est-il de l'ensemble des configurations ? Si l'on reprend le plan factoriel de la figure 12 pour y projeter les traits des intertitres du corpus, une nouvelle opposition se dessine nettement, confirmant le lien entre type d'intertitre et genre textuel. La figure 13 (reprise de Jacques & Rebeyrolle, 2006) [18] montre un regroupement des intertitres des articles scientifiques dans la zone que nous avons caractérisée comme celle de l'implication thématique alors que les intertitres des textes professionnels se placent dans la zone de l'implication référentielle. Et effectivement, dans le cadre de la rédaction de cette synthèse, j'ai cherché des exemples d'intertitres référentiels dans le corpus Termith et n'en ai guère trouvé. L'écriture scientifique procède plus par structuration thématique que par focalisation sur un/des référents.

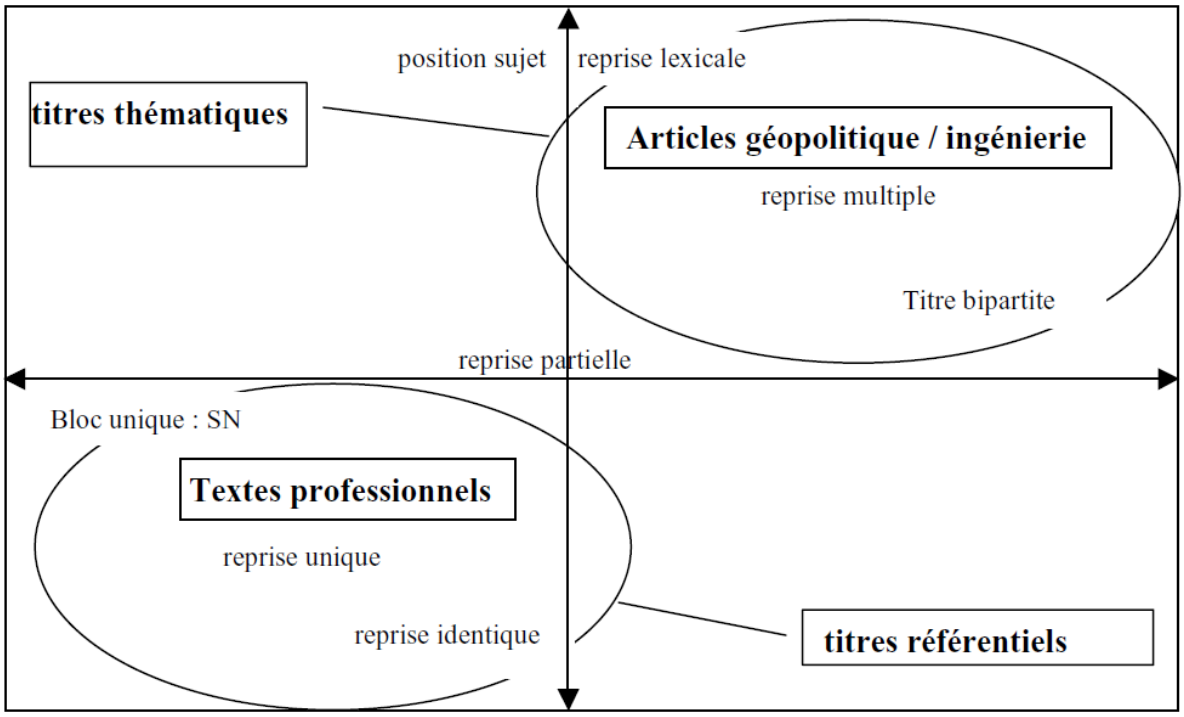


Figure 13 : *Projection des sous-corpus sur les axes dégagés par l'analyse multifonctionnelle*

Le contraste avec un genre bien différent, la presse écrite, complète les observations.

4.2.2.5 Intertitres et titres de presse : deux modalités d'analyse nécessairement différentes

C'est un trio un peu différent qui a poursuivi l'analyse des fonctions des titres dans (Rebeyrolle, Jacques & Péry-Woodley, 2009) [20]. Il s'agissait plutôt de mettre en évidence, par le contraste avec un corpus de titres de faits divers et son analyse, les spécificités d'une démarche et l'enrichissement qu'elle apporte aux études sur le fonctionnement et l'organisation du discours.

Le fait de mettre en regard l'analyse de titres de faits divers et « nos » intertitres de documents longs et structurés pourrait paraitre une approche biaisée : peut-on réellement comparer des titres de faits divers, conçus pour être « accrocheurs », par définition très contextualisés culturellement et temporellement, intitulant des textes relativement brefs, avec les intertitres qui découpent et structurent des documents longs, en texte spécialisé, donc en général plus

73 Raballand G. (2003). Géoéconomie du bassin Caspien. IFRI.

soucieux de clarté et de précision que d'effets rhétoriques ? C'est précisément cette différence entre les deux sortes de titres que nous avons voulu faire apparaitre, pour en conclure à la nécessité d'un « appareillage » adapté à chacun mais aussi en faire émerger des mutualisations possibles en termes de démarche. L'analyse de ces deux types de titres présente tout de même certains points communs : dans les deux cas, la relation entre le titre et le texte qu'il intitule est interrogée, la question du traitement cognitif « préconisé » – implicitement, s'entend – par le titre est envisagée.

En ce qui concerne les titres de faits divers, la question principale était celle du modèle le plus pertinent à mobiliser pour rendre compte de la relation entre le titre et le texte du fait divers. Dans l'hypothèse où le lecteur d'un fait divers intègre titre et texte dans une même séquence discursive complexe au cours de l'interprétation, quelle est la nature de leur articulation ? Bien que – comme je l'ai signalé en début de chapitre – les titres de presse aient fait déjà l'objet d'analyses linguistiques, cette question n'a guère été traitée. Et des divers modèles qui proposent une théorie des relations de discours, seule la *Rhetorical Structure Theory* – Théorie des Structures Rhétoriques – (RST, Mann & Thompson, 1988) prend explicitement en compte le rôle de tels titres et propose de les voir comme un satellite du texte, relié à celui-ci par la relation *Preparation*. Sur le plan cognitif, une telle relation signifie que le matériau apporté par le titre permet au lecteur de se préparer cognitivement à ce qui va suivre : anticiper, mobiliser des connaissances d'arrière-plan, mettre un référent qui serait mentionné dans le titre au premier plan, etc. L'exemple (60) en est une bonne illustration : à partir du titre, le lecteur dispose déjà de l'information qui va être développée – des non-lieux –, il est invité à placer au centre de son attention les deux référents *Jérôme Kerviel* et *la Société générale* et à retrouver dans sa mémoire les éléments de contexte pour ces deux référents, c'est-à-dire l'affaire financière qui a causé des pertes conséquentes à la banque et qui a opposé celle-ci au *trader* Jérôme Kerviel.

(60) Jérôme Kerviel contre la Société générale : deux non-lieux
Selon les juges d'instruction, la Société générale n'a pas cherché à manipuler la justice dans l'affaire Kerviel, suite aux plaintes de l'ex-trader pour "escroquerie au jugement" et "faux et usage de faux" déposées par l'ancien trader contre la banque.
Conformément aux réquisitions, deux non-lieux ont été prononcés dans les enquêtes où Jérôme Kerviel accusait la Société générale de manipulation. [extrait du Dauphiné Libéré du 4/08/2017]

Dans ce que je viens d'évoquer, on retrouve certaines des caractéristiques des intertitres à implication thématique, en particulier cette notion d'anticipation, de préparation. Le modèle de la RST peut ainsi sembler adéquat pour l'analyse aussi bien des titres de faits divers que des intertitres.

La différence majeure qui a justifié de ne pas analyser les intertitres en termes de relations de discours telles que celles qui sont proposées par la RST réside dans la participation de l'intertitre à la structuration globale d'un texte long, qui implique des relations plus complexes que celle qui lie l'intertitre à ce qu'il titre : relations avec l'amont du titre, relations avec les autres titres, notion de portée, ainsi que je l'ai développé précédemment. C'est cette complexité qui a motivé l'élaboration de la démarche de type *corpus-driven* (Tognini-Bonelli, 2001) présentée ci-dessus (section 4.2.1) : il ne s'agissait pas de projeter un modèle théorique à la lumière duquel examiner les données mais de partir du corpus lui-même pour en faire émerger des fonctionnements réguliers.

Il ressort du contraste entre ces types différents de titres et entre les deux approches qui ont été mises en œuvre pour en cerner le fonctionnement (codage de traits formels *vs* projection de relations de discours*)* que trois aspects reviennent de façon cruciale dans la caractérisation des fonctionnements discursifs :

1. le relief particulier du titre, de par sa singularité matérielle, le contraste qu'il présente avec le corps du texte, l'emploi d'une mise en forme à fort contraste (souvent de l'italique, du gras, une police de caractères spécifique), il est un lieu visiblement saillant du texte, propriété qui s'étend à la proposition ou au syntagme qu'il exprime, comme nous allons le voir ensuite (section 4.3.1) ;
2. le rôle instructionnel du titre, plutôt tourné vers la gestion de la référence ou plutôt tourné vers celle de la thématique, vérifié par les deux types de titres (presse et intertitres) ;
3. la dimension structurante de l'intertitre, par sa capacité de segmentation / regroupement de sections dans des documents longs et par sa capacité à construire une hiérarchie.

Notre approche descriptive des titres et intertitres, visant à en cerner les fonctions en nous appuyant sur des traits formels, est de celles qui sont mises à profit par les recherches de TAL (Traitement Automatique des Langues) qui portent sur la génération de texte et en particulier sur le titrage automatique (Lopez, 2013). L'enjeu pour ces recherches est de réaliser automatiquement des textes ou des résumés qui ne paraissent pas artificiels. Une compréhension de la façon dont les intertitres agissent dans les textes permet donc d'améliorer les traitements automatiques, non seulement en analyse, mais en génération.

J'ai poursuivi ces travaux sur les intertitres, de façon plus personnelle, dans le (nouveau) cadre de travail qu'est devenu pour moi le Lidilem.

4.3 Les intertitres et la rhétorique de l'article scientifique

Le cadre offert par le Lidilem s'est avéré propice à une focalisation sur les textes scientifiques, objets privilégiés du projet Scientext et du projet Termith, que j'ai évoqués dans le chapitre 3. Ce n'est plus toutefois sous l'angle de la découverte des fonctions discursives des intertitres que je les ai regardés dans ces nouveaux corpus, mais en approfondissant la problématique de la mise en relief ou avec l'éclairage des recherches sur la rhétorique propre à l'écrit scientifique.

4.3.1 L'intertitre comme lieu de mise en relief

Si l'intertitre joue un tel rôle instructionnel, c'est parce qu'il se détache de façon visible dans la matérialité du texte. Le contraste de mise en forme qui le distingue et lui permet de fonctionner lui octroie une saillance particulière, dont bénéficient les éléments syntaxiques qui le composent, de quelque nature qu'ils soient. C'est ce que j'ai développé dans (Jacques, 2017a) [33].

À partir de la double constatation que, comme je l'ai déjà indiqué, un texte long doit jouer sur les capacités mémorielles du lecteur en fournissant des moyens de remettre « sur le devant de la scène » certaines entités pour pouvoir les manipuler dans le discours, et que les travaux sur la saillance tels que ceux de Landragin (2011, 2011, 2012) posent le contraste comme facteur de saillance, j'ai pris appui sur les travaux exposés dans la section 4.2 pour examiner plus particulièrement la façon dont les intertitres « travaillent » la mise en relief. L'idée sous-jacente est qu'ils concourent à la dynamique de la gestion des entités du texte – terme délibérément vague afin d'englober aussi bien les référents, procès, évènements ou autres que le texte est

amené à « mettre en scène » – et ce, de par précisément le double fait qu'ils introduisent une rupture dans le texte, en opérant une segmentation, et qu'ils accroissent la visibilité de leurs constituants.

En quoi consiste cette mise en relief, quelles opérations réalise-t-elle ? En sélectionnant dans les divers corpus présentés jusqu'ici les textes relevant du discours scientifique, j'en ai distingué deux principales.

1. Activation d'un référent

Dans les travaux qui précèdent, pour cerner les fonctions des intertitres, nous étions parties du matériel lexical de l'intertitre et avions examiné son devenir dans la section. La perspective est ici différente puisqu'elle consiste à regarder ce qui se passe immédiatement après l'intertitre.

Les travaux sur l'accessibilité cognitive des référents (de Mulder, 2000 ; Gundel, 1998 ; Gundel *et al.*, 2000) mettent en rapport la forme de la désignation d'un référent et son accessibilité : moins la forme est spécifiée, plus le référent est réputé accessible – d'où des échelles d'accessibilité qui attribuent un indice d'accessibilité plus grand aux expressions linguistiques les plus sous-spécifiées telles que les pronoms[74]. Donc l'emploi d'expressions linguistiques peu spécifiées après un intertitre, pour en opérer une anaphore, signale que le référent ainsi désigné est (supposé) très accessible. J'en déduis qu'il est considéré comme saillant par l'auteur du texte (à raison ou à tort). Plusieurs expressions manifestent cette accessibilité élevée : un déterminant démonstratif (61), le pronom démonstratif « ce » (62) ou un pronom personnel (63).

(61) 2.2 La culture de l'information orientée « bibliothèque »
Cette conception est clairement celle qui domine au niveau international, avec notamment les actions menées en ce qui concerne l'information literacy. Même si ce concept anglo-saxon n'est pas tout à fait équivalent à celui de culture de l'information, il existe de fortes proximités. [Termith - sciences de l'information et de la communication]

(62) 3.3. La troncation
C'est l'une des stratégies les plus courantes de la manipulation lexicale notamment pour les langues mixtes, les langues de jeunes et les argots. [Termith - linguistique]

(63) Surveillance d'un syndrome HNPCC
Elle s'applique (cas index et apparentés) : aux personnes porteuses d'une mutation constitutionnelle d'un gène MMR […], à celles en attente d'un résultat de recherche d'une mutation constitutionnelle, à celles ayant refusé soit la consultation d'oncogénétique [Médecine[75]]

J'ai vérifié dans chacun des cas que le référent visé par l'expression anaphorique n'avait pas été introduit dans le discours immédiatement avant l'intertitre, c'est donc celui-ci qui lui confère cette saillance. Celle-ci peut se maintenir tout au long d'une section, ce qui témoigne de la seconde opération de mise en relief assurée par l'intertitre.

74 À propos de l'accessibilité, voir la note 8, page 21.

75 Il s'agit ici du corpus de textes de recommandations médicales réuni pour l'analyse de l'expression des procédures (cf. 2.2.2).

2. Portée textuelle étendue

L'intertitre confère à son contenu une portée textuelle qui s'étend sur l'ensemble de la section titrée et des sous-sections. En témoigne l'extrait (64), dans lequel j'ai dû opérer des coupes afin d'en restreindre la longueur. Pour la lisibilité, les intertitres sont en gras et italique pour le niveau le plus haut, en italique pour les niveaux sous-ordonnés.

(64) ***Le traitement médical et chirurgical des kystes présumés fonctionnels***

Les kystes fonctionnels spontanés

Chez la femme en période d'activité génitale, le traitement médical progestatif macrodosé, (danazol, oestroprogestatifs) n'est pas recommandé [...].

Chez la femme ménopausée, en l'absence de facteurs de risque et de symptomatologie clinique, l'abstention thérapeutique est recommandée [...].

Les kystes fonctionnels induits par les traitements médicaux

[…] Avec les stérilets au levonorgestrel, des kystes ovariens fonctionnels sont observés [...]. Il n'y a pas lieu de prescrire un traitement médical ni de les ponctionner (NP5).

Les KOF induits par des traitements inducteurs de l'ovulation paucifolliculaire disparaissent spontanément avec l'arrêt du traitement. Il faut recommander l'abstention thérapeutique (NP2).

[…]

Les KOF induits par les traitements chirurgicaux

En l'absence de symptomatologie ovarienne, le traitement médical ou chirurgical de la dystrophie kystique des ovaires sous-adhérentiels n'est pas recommandé (NP5). […] [Médecine]

Quoique la section intitulée *Le traitement médical et chirurgical des kystes présumés fonctionnels* soit découpée en trois sous-sections, le texte de l'ensemble de la section continue d'évoquer le traitement médical et chirurgical pour chacune des pathologies indiquées dans les intertitres de niveau inférieur. Chacune des sous-sections est en fait à propos non des *kystes fonctionnels spontanés*, des *kystes fonctionnels induits par les traitements médicaux*, des *kystes ovariens fonctionnels induits par les traitements chirurgicaux* mais du **traitement médical et chirurgical** de tous ces *kystes fonctionnels*. Une répétition systématique aurait considérablement alourdi le texte, l'intertitre offre ici le moyen d'opérer une certaine factorisation. Celle-ci atteint une forme encore plus accomplie si l'intertitre se complète syntaxiquement sur les intertitres de niveau inférieur, extrait (65).

(65) ***IV. Surveillance à long terme***

1. De la mère

Le diabète gestationnel constitue un marqueur précoce du risque de survenue d'un diabète non insulino-dépendant. Il est donc nécessaire de surveiller régulièrement la tolérance au glucose de ces femmes […] Un dépistage et un traitement précoces doivent être assurés en cas de nouvelles grossesses (NP2).

2. De l'enfant

Le risque d'obésité et de diabète non insulino-dépendant est augmenté chez ces enfants. Une surveillance régulière des enfants et une éducation nutritionnelle de la mère et de l'enfant sont donc nécessaires. [Médecine]

L'intertitre de niveau supérieur fournit la tête d'un groupe nominal qui n'est complet qu'avec les intertitres de niveau inférieur.

Ces divers exemples témoignent du double apport des intertitres dans la construction textuelle. À un niveau local, il assure saillance et accessibilité à son contenu, à un niveau plus global, il structure le propos sur une portée étendue en permettant une certaine économie de moyens

dans la mise en place de relations entre les sections, et entre les sections et le texte dans son ensemble.

C'est le niveau du texte dans son ensemble que j'ai examiné selon deux nouvelles perspectives que j'aborde maintenant : la structure de l'article scientifique, la comparaison de la structuration à l'oral et à l'écrit.

4.3.2 Les intertitres et la structuration de l'article scientifique

Le projet Scientext[76] a offert à la communauté scientifique une base textuelle qui constitue un observatoire de choix pour la compréhension de l'écriture scientifique. J'y ai donc sélectionné la matière d'une nouvelle orientation de recherche, complémentaire aux travaux précédents (voir aussi Hartwell & Jacques, 2014). Il s'agissait de s'intéresser d'encore plus près à la dimension idéationnelle et discursive de l'intertitre en regardant comment, dans les articles de sciences humaines et sociales, il structure la logique de l'exposé scientifique [27].

Le point de départ de la réflexion est le constat d'une structure récurrente dans les sciences expérimentales, le format IMRAD (*Introduction, Methods, Results, And Discussion*) qui s'est constitué et peu à peu imposé depuis la médecine. Il est instructif de noter que, dans leur enquête pour établir la période à laquelle ce format est devenu le standard dans les articles de recherche en médecine, Sollaci et Pereira (2004) se basent sur les titres de section (*headings*) pour décider si un article se conforme ou ne se conforme pas au format IMRAD :

> An article was considered to be written using the IMRAD structure only when the headings "methods, results, and discussion," or synonyms for these headings, were all included and clearly printed. The introduction section had to be present but not necessarily accompanied by a heading. Articles that did not follow this structure were considered non-IMRAD. (Sollaci & Pereira, 2004)

C'est donc avant tout le découpage et l'intitulé des sections qui signalent le respect de la structure, plus que le contenu lui-même. Et ceci, sans doute parce que les intertitres jouent ici un rôle de formulaire, pré-organisant la logique de la démonstration :

> avant même de commencer à renseigner les différentes sections de l'article, la logique argumentative inscrite matériellement dans les titres dicte un mode d'énonciation séquencé bien défini. (Pontille, 2007)

Ce format a « essaimé » dans les diverses sciences expérimentales, sans doute, pensent Sollaci et Pereira (2004), en raison de la facilitation de la lecture et de l'évaluation qu'il offre. Mais il ne s'impose pas dans les SHS. La question est alors de déterminer comment les articles de SHS gèrent cette absence de prescription explicite de structure : chaque auteur scientifique fait-il/elle « comme il/elle veut » ? Ou y a-t-il au contraire une structure implicite ? Une même structure « des SHS » ou une structure propre à chaque discipline ?

J'ai mis à profit le corpus Scientext et celui de Termith pour une analyse contrastive d'articles de sociologie, de sciences de l'information et de la communication et de traitement automatique des langues. J'ai sélectionné une trentaine d'articles (environ 10 de chaque discipline), soit en tout 421 intertitres. Je reprends de (Jacques, 2014 : 204) [27] le tableau 5 indiquant la répartition :

76 http://scientext.msh-alpes.fr/scientext-site/spip.php?article1

Discipline	Nombre d'articles	Nombre d'intertitres
Traitement automatique des langues (TAL)	8	151
Sociologie (SOC)	10	120
Sciences de l'information et de la communication (SIC)	10	150
Total	**28**	**421**

Tableau 5 : *Corpus d'étude*

1. Premier axe d'analyse : la forme des intertitres

Pour analyser ces intertitres, j'ai repris entre autres la dichotomie entre intertitres formels et intertitres idéationnels. Je rappelle que, pour le MAT, les premiers sont ceux qui entretiennent une relation d'identification avec « leur » section ; par exemple *Introduction* ou *Conclusion* : la section ainsi intitulée **est une** conclusion ou une introduction mais n'est pas **à propos** d'une introduction ou d'une conclusion. Les intertitres hybrides associent un contenu plus idéationnel à ce signalement formel, par exemple « *Conclusion. Des difficultés inhérentes à la Chine ?* » (dans un article de sciences de l'information et de la communication).

L'emploi de tels intertitres, formels et hybrides, est un facteur de différenciation des disciplines : ils représentent moins de 10 % des intertitres de sociologie et 15 % à 16 % de ceux des deux autres disciplines. La sociologie semble donc se démarquer fortement, notamment par le fait qu'aucun des 10 articles analysés ne comporte d'intertitre *Introduction* : les titres formels relevés y sont seulement *Conclusion* et *Bibliographie*. Mais peut-être est-ce dû au caractère ressenti comme scolaire de ces intertitres formels : de fait, un auteur scientifique n'a en général pas besoin de signaler le début de son article comme étant une introduction, le simple fait qu'une section se situe au début du texte la catégorise implicitement comme introduction.

Cette première opposition, TAL et SIC d'un côté, SOC de l'autre, est mise à mal lorsque l'on regarde la forme grammaticale des intertitres. Je reprends le tableau 6 de (Jacques, 2014 : 206) [27] :

	TAL	**SIC**	**SOC**
Syntagme Nominal (SN)	125 (94,7%)	111 (86,7%)	92 (84,4%)
Syntagme Verbal (SV)	7 (5,3%)	13 (10,2%)	5 (4,6%)
Syntagme Prépositionnel (SP)	0	3 (2,3%)	8 (7,3%)
Phrase (Ph)	0	1 (0,8%)	4 (3,7%)
Total	132 (100%)	128 (100%)	109 (100%)

Tableau 6 : *Formes grammaticales des intertitres selon la discipline*

Tous les travaux que j'ai évoqués jusqu'ici dans ce chapitre ont constaté la prédominance du SN parmi les intertitres, elle se confirme ici. Elle est une quasi-exclusivité pour les auteurs de TAL – surtout quand on sait que 6 des 7 SV relevés appartiennent au même article[77], par

77 Antoine J.-Y., Maurel D. (2007). Aide à la communication pour personnes handicapées et prédiction de texte. *TAL* 48(2), pp. 9-46.

exemple *3. Accélérer la saisie : sélection rapide sur le clavier virtuel*. Se dessine donc ici une opposition différente : SIC et SOC d'un côté, TAL de l'autre.

La forme grammaticale des intertitres est un des paramètres qui interviennent sur le contenu. On a vu précédemment que le pôle d'implication référentielle se caractérise entre autres par la forme SN « monobloc ». La diversification croissante, de TAL à SIC, des formes grammaticales laisse présager des opérations discursives plus variées.

2. Logique discursive inscrite dans les intertitres

Comme le format IMRAD constitue un découpage de premier niveau des articles, j'ai examiné les intertitres de niveau 1, pour rester au même niveau de structuration. La conclusion est simple, ces trois disciplines, d'une part, entretiennent plus ou moins de proximité avec IMRAD, d'autre part, organisent la logique de l'exposé scientifique non de façon uniforme mais en fonction des thématiques et des objets travaillés.

La proximité avec IMRAD a été appréciée en fonction de l'adoption de l'une des sections-types de IMRAD : *méthode, résultats, discussion* (j'ai indiqué ce qu'il en était de l'introduction ci-dessus). La sociologie n'en emploie aucun, le TAL et les SIC y ont ponctuellement recours : 3 intertitres sur 151 pour le TAL, 7 sur 150 pour les SIC. Cette disparité ne signifie aucunement que la sociologie ne présente ni ses méthodes ni ses résultats et ne les discute pas, mais qu'elle ne les affiche pas en tant qu'organisateurs de l'article scientifique. J'y reviens plus loin.

Chacune des trois disciplines présente une organisation des articles cohérente avec le type d'objets travaillés par la discipline. Quand le TAL se préoccupe de mettre au point des outils informatiques reliés à des tâches langagières, les articles commencent par expliciter la problématique, indiquer un état de l'art qui se conclut généralement sur les lacunes ou les verrous des solutions jusque-là envisagées et poursuivent en explicitant la solution proposée par les auteurs, solution ensuite évaluée dans une section dédiée. La série d'intertitres de niveau 1 reproduite en (66) est un exemple parfait de cette structuration.

(66) 1. Introduction : communication assistée et prédiction de texte
2. Aide à la communication pour personnes handicapées
3. Accélérer la saisie : sélection rapide sur le clavier virtuel
5. Communiquer sans langage verbal
6. Évaluation
7. Conclusion[78] [Scientext - TAL]

Quand il s'agit de bâtir ou d'évaluer des formalismes, des modélisations, les intertitres donnent plutôt une impression d'énumération : les différents formalismes sont « passés en revue » de façon systématique, ou bien le problème à traiter est éclairé successivement sous ses différentes facettes. Là encore, la série d'intertitres de niveau 1 reproduite en (67) montre cette structuration : l'examen successif de tous les modèles de discours annoncés dans le titre de l'article « *Capacité générative forte de RST, SDRT et des DAG de dépendances pour le discours* », puis l'évaluation de leur « capacité générative ».

(67) 1. Introduction
2. RST
3. SDRT
4. DAG de dépendances pour le discours

78 Antoine J.-Y., Maurel D. (2007). Aide à la communication pour personnes handicapées et prédiction de texte. *TAL* 48(2), pp. 9-46.

5. Capacité générative forte
6. Résumé et conclusion[79] [Scientext - TAL]

Cette dernière structuration se retrouve dans les articles de sciences de l'information et de la communication, discipline qui est en fait à la croisée des deux autres avec schématiquement des recherches plus sociologiques et d'autres plus informatiques. Le corpus met en évidence deux types d'articles de recherche pour cette discipline : soit une enquête de terrain, pour laquelle sont explicités la méthodologie et les résultats, ce qui conduit à une structuration très proche de IMRAD, comme en (68), soit une analyse d'un phénomène, auquel cas on retrouve cette impression de passage en revue systématique de toutes les facettes de la question, comme le montre bien (69).

(68) Introduction
Positionnement théorique
Méthodologie et terrains mobilisés
Résultats
Conclusion[80] [Termith – sciences de l'information et de la communication]

(69) Partager l'expérience de visionnage
Partager sa passion à travers les conversations
Visionnage et partage, deux activités entrelacées
Partager l'accès au contenu
Conclusion[81] [Termith – sciences de l'information et de la communication]

On se rapproche en (69) tout à fait de la démarche et donc de la structure de l'article de sociologie qui, typiquement, bâtit une argumentation progressive autour de la question de recherche traitée. Les intertitres sont autant de propositions réduites qui affichent soit les thèses qui vont être défendues par l'auteur, soit les questions qui organisent son argumentation, voir (70).

(70) L'éclectisme des goûts, nouvel horizon de la légitimité culturelle ?
Asymétrie des échanges symboliques
Un relâchement de l'affinité entre musique classique et classes supérieures ?
Les fondements structurels de l'« omnivorité » culturelle
Goûts de classe et culture de masse[82] [Termith – sociologie]

J'ai été frappée par la récurrence d'intertitres de la forme « X comme Y » qui mettent en scène non une comparaison mais une identification, (71) et (72) en présentent un aperçu.

(71) L'implicite persistant de la localité comme générateur de relation sociale
L'émergence d'un voisinage rural comme espace de « repos social »[83] [Termith – sociologie]

79 Danlos L. (2006). Capacité générative forte de RST, SDRT et des DAG de dépendances pour le discours. *TAL* 47(2), pp. 169-198.

80 Vacher B. (2010). Sens et normes font-ils bon ménage dans les organisations ? *Études de communication* 34, pp. 127-142.

81 Combes, C. (2011). La consommation de séries à l'épreuve d'internet: Entre pratique individuelle et activité collective. *Réseaux* 165(1), pp. 137-163.

82 Coulangeon, P. (2010). Les métamorphoses de la légitimité : Classes sociales et goût musical en France, 1973-2008. *Actes de la recherche en sciences sociales* 181-182(1), pp. 88-105.

83 Banos, V., Candau, J. & Baud, A. (2009). Anonymat en localité : Enquête sur les relations de voisinage en milieu rural. *Cahiers internationaux de sociologie* 127(2), pp. 247-267.

(72) Le marché intérieur comme recomposition des intérêts économiques et sociaux
Le dialogue social européen comme ancrage communautaire d'une activité législative[84] [Termith – sociologie]

Ce *comme* éclaire la perspective selon laquelle le/la chercheur/e analyse et présente le phénomène qu'il/elle étudie dans l'article. C'est lui/elle qui attribue le second terme (après *comme*) au premier (avant *comme*). Le « travail » de la section est alors de montrer cette homologation de X à Y.

Il ne faudrait pas en conclure que la sociologie ou les articles de sciences de l'information et de la communication qui manifestent les mêmes caractéristiques de structuration que la sociologie font totalement l'impasse sur la méthodologie et sur les données. Ces éléments sont bien présents dans les articles mais sont, ou dilués dans le texte, ou, pour la sociologie[85], traités dans des encadrés.

Cette étude montre que hors du champ des sciences expérimentales, là où IMRAD n'est pas explicitement prescrit, la structure de l'article telle que livrée par les intertitres dépend indirectement des disciplines, elle est en fait davantage liée aux objets travaillés et à la mise en scène des apports du chercheur : plus ces apports sont fondés sur une construction intellectuelle, une déduction basée sur les données empiriques recueillies par le chercheur par observation ou enquête, moins l'article fait usage des rubriques du format IMRAD. Celles-ci se révèlent plus appropriées aux démarches expérimentales fondées sur la manipulation et l'évaluation des résultats de la manipulation. Plus le travail du chercheur consiste à proposer un système interprétatif des données recueillies, plus il construit les intertitres comme autant de propositions dont la validité est étayée par le texte de la section – le parangon en l'occurrence étant la sociologie, qui n'a aucun moyen de manipuler des variables et d'agir sur les structures sociales mais vise à en décrire la logique sous-jacente. Les intertitres jouent alors un rôle rhétorique majeur.

Qu'en est-il lorsque l'exposé scientifique n'est plus écrit mais oral ?

4.4 À l'oral, que deviennent les fonctions des intertitres ?

Cette dernière étude a eu pour cadre le projet EIIDA, Étude Interdisciplinaire et Interlinguistique du Discours Académique (Carter-Thomas & Jacques, 2017) [31], piloté par Shirley Carter-Thomas. Le projet s'est donné pour objectif « de comparer le discours scientifique écrit et le discours scientifique oral, et d'interroger l'impact de la transmission directe sur le discours scientifique »[86], et ce dans trois langues différentes : anglais, espagnol, français. Un corpus comparable[87] a été réuni pour cette étude. Il est comparable sous trois versants. Il rassemble un matériau similaire, des exposés scientifiques :

1. sous forme écrite et sous forme orale, articles de recherche et présentations de conférences ;

2. dans les trois langues, anglais, français, espagnol ;

3. dans deux disciplines de champs contrastés, la linguistique et la géochimie organique.

84 Didry, C. (2009). L'émergence du dialogue social en Europe : retour sur une innovation institutionnelle méconnue. *L'Année sociologique* 59(2), pp. 417-447.

85 Les articles de sociologie sont les seuls du corpus à juxtaposer des encadrés au corps du texte, voir par exemple Caveng, R. (2009). Inversement des positions et ré-enchantement de l'interaction: La relation d'enquête dans les sondages et les études de marchés. *Actes de la recherche en sciences sociales* 178(3), pp. 88-97.

86 http://lattice.cnrs.fr/Projet-Etude-interdisciplinaire-et?lang=fr page consultée le 5 aout 2017.

87 Malheureusement, malgré nos efforts, ce corpus n'est pas totalement libre de droits.

La figure 14 en précise la composition en indiquant pour chaque article de recherche et pour chaque présentation de conférence le nombre de textes réunis et le nombre de mots de chaque sous-corpus.

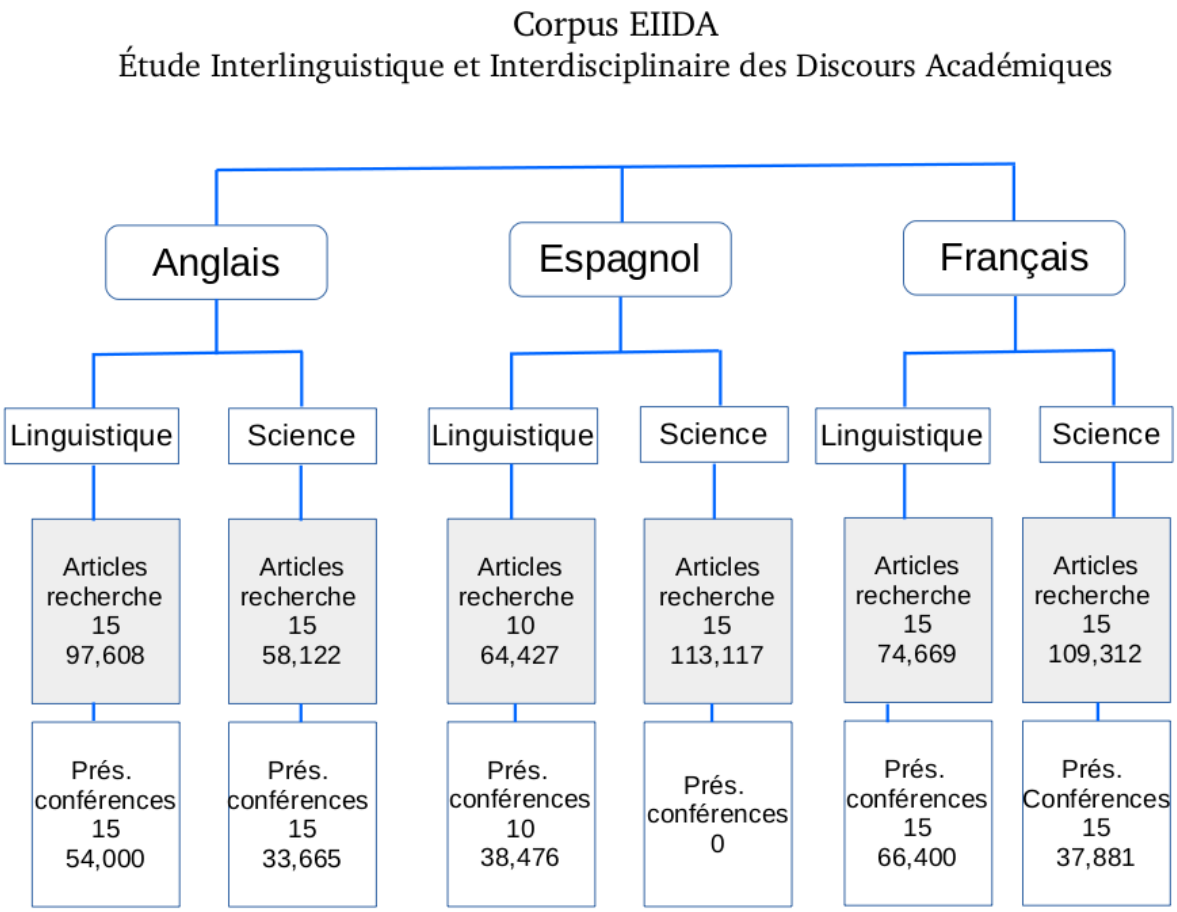


Figure 14 : *Structure, quantité de textes et de mots du corpus EIIDA*

Une telle composition se destinait à être le support d'études contrastives variées, dont une partie a été réunie dans un numéro de la revue Chimera[88], et m'a permis d'interroger les moyens de structuration de l'exposé scientifique à l'oral (Jacques, 2017b) [32], en prenant en compte le fait que les présentations de conférence associent un support visuel, un diaporama, à l'exposé oral, ce qui contribue à la gestion de la présentation (Rowley-Jolivet & Carter-Thomas, 2005). C'est vers le corpus français en géochimie organique que je me suis tournée, car dans cette discipline, pour l'oral nous disposions du film des conférences donc du « texte » audio en même temps que du diaporama projeté[89]. La partie vidéo se limite au diaporama et n'inclut pas l'orateur. Les exposés ont été transcrits et transformés en xml. J'ai ajouté dans chaque fichier un repérage des changements de diapositive pour les raisons de mon étude.

Ce matériau permet d'explorer les questions suivantes : la structuration réalisée dans un article de recherche par les intertitres a-t-elle une contrepartie à l'oral ? Comment l'auditeur est-il aidé, guidé, pour se repérer dans la structure de la présentation de conférence ? Ce guidage est-

88 https://revistas.uam.es/index.php/chimera/issue/view/679 page consultée le 5 aout 2017.

89 Le programme de la conférence est visible ici http://geochimie.fr/reunions-annuelles-en-geochimie-organique/reunion-annuelle-des-geochimistes-organiciens-francais-de-2012/
Les vidéos sont disponibles sur la « chaine » de Jérémy Jacob http://www.dailymotion.com/Jeremy-Jacob/videos pages consultées le 5 aout 2017.

il assuré par le visuel qui accompagne la présentation, le diaporama ? Mais la première de toutes les questions étant à vrai dire la nécessité et la pertinence d'un guidage. En effet, si articles et présentations de conférence dans ce domaine se conforment à une structuration préétablie, le format IMRAD qui organise volontiers les sciences expérimentales, le guidage de l'auditeur peut être réduit et allégé.

La première analyse menée sur ce corpus est donc la vérification de la présence du format IMRAD dans les articles de recherche, avec le critère simple mis en œuvre par (Sollaci & Pereira, 2004): la mention des « rubriques » IMRAD dans les intertitres. En fonction de ce critère, aucun article ne se conforme strictement à IMRAD. Le tableau 7 comptabilise, pour chaque rubrique IMRAD, le nombre d'articles qui l'adoptent comme intertitre de niveau 1 et, entre parenthèses, comme intertitre de niveau 2. On lit ainsi sous *Méthode* que 9 articles présentent une telle section au niveau 1 et 1 article intitule ainsi une section de niveau 2 ; *Résultats* est présent comme section de niveau 1 dans 5 articles et comme section de niveau 2 dans 2 articles.

Introduction	**Matériel**	**Méthode**	**Résultats**	**Discussion**
15	6	9 (1)	5 (2)	6 (1)

Tableau 7 : *Présence des « rubriques » IMRAD dans les articles de recherche de géochimie organique du corpus EIIDA*

Le format IMRAD n'est donc pas absent mais il est aménagé, pour laisser une place à une section qui explicite le contexte géographique de l'étude et qui est titrée, selon les cas, « *2. Le cadre géographique* » « *2. Contexte géographique* » « *2. Cadre géoarchéologique* » « *2. Contexte géographique et écologique* » « *2. Cadre topographique et géologique* » « *2. Site d'étude* » « *2. Le secteur d'étude* ».

Le schéma d'organisation qui se dessine alors adopte la structure suivante :

- Introduction
- Contexte / cadre géographique
- Matériel / sources
- Méthode / données
- Résultats
- Discussion / Interprétation
- Conclusion

Les intertitres reproduits en (73), (74) et (75) illustrent des réalisations de ce schéma et montrent les libertés que les articles de cette discipline prennent avec le format IMRAD.

(73) 1 – INTRODUCTION
2 - LE BASSIN VERSANT ET LA ZONE HUMIDE DE MONTCHÂTRE
3 – MÉTHODE
4 – RÉSULTATS
5 – DISCUSSION
6 – CONCLUSION[90] [EIIDA - géochimie - écrit]

(74) 1 – INTRODUCTION
2 - CONTEXTE GÉOGRAPHIQUE
3 - POSITION SÉDIMENTAIRE DES CAROTTAGES ÉTUDIÉS
4 - MÉTHODES ET MATÉRIAUX

90 Ballut C., Prat B., López-Sáez J. A., Gaby G. et Cabanis M. (2008). Evolution environnementale d'une zone humide et de son bassin versant depuis la fin de l'âge du fer, *Quaternaire* 19/1, pp. 69-79.

5 - RÉSULTATS SUR LES PHASES SÉDIMENTAIRES, ALGOLOGIQUES ET LES VARIATIONS DU MILIEU LIMNIQUE
6 - RÉSULTATS SUR LES TRANSFORMATIONS DU BASSIN VERSANT PAR L'HOMME
7 - DISCUSSION SUR LES VARIATIONS DU LAC DE PALADRU ET DANS LES MILIEUX CONTINENTAUX VOISINS
8 - DISCUSSION SUR LES DEUX PÉRIODES D'OPTIMUM CLIMATIQUE DE L'ANTIQUITÉ ET DU MOYEN ÂGE
9 – CONCLUSIONS[91] [EIIDA - géochimie - écrit]

(75) 1 – INTRODUCTION
2 - CADRE TOPOGRAPHIQUE ET GÉOLOGIQUE
3 - CHRONOSTRATIGRAPHIE ET SÉDIMENTOLOGIE DES DÉPÔTS LACUSTRES
4 - MÉTHODE DE CALCUL DES TAUX D'ÉROSION
5 - INTERPRÉTATION DES RÉSULTATS
6 - CONCLUSION[92] [EIIDA - géochimie - écrit]

Si la structure de l'exposé scientifique écrit dans la discipline n'est pas complètement fixée à l'avance et supporte une marge de variation, il est probable que l'exposé oral en fasse de même. Et puisqu'il y a variation potentielle, on peut s'attendre à un guidage de l'auditeur par rapport à cette structuration qui n'est pas totalement attendue.

Pour apprécier la structuration des exposés oraux, je me suis fiée aux intitulés présents sur les diapositives projetées et en particulier à la présence d'un « bandeau » qui précise à la fois la structure globale de l'exposé et le point de l'exposé de la diapositive courante. La figure 15 donne un exemple d'un tel bandeau dans le contexte de sa diapositive.

91 Brochier J.-L., Borel J.-L., Druart J.-C. (2007). Les variations paléoenvironnementales de 1000 avant à 1000 après J.C. et la question des « optima » climatiques de l'Antiquité tardive et du Moyen Âge sur le piémont des Alpes du nord à Colletière, lac de Paladru, France, *Quaternaire* 18/3, pp. 253-270.

92 Degeai J.-P., Pastre J.-F. (2009). Impacts environnementaux sur l'érosion des sols au Pléistocène supérieur et à L'holocène dans le cratère de maar du lac du Bouchet (Massif central, France), *Quaternaire* 20/2, pp. 149-159.

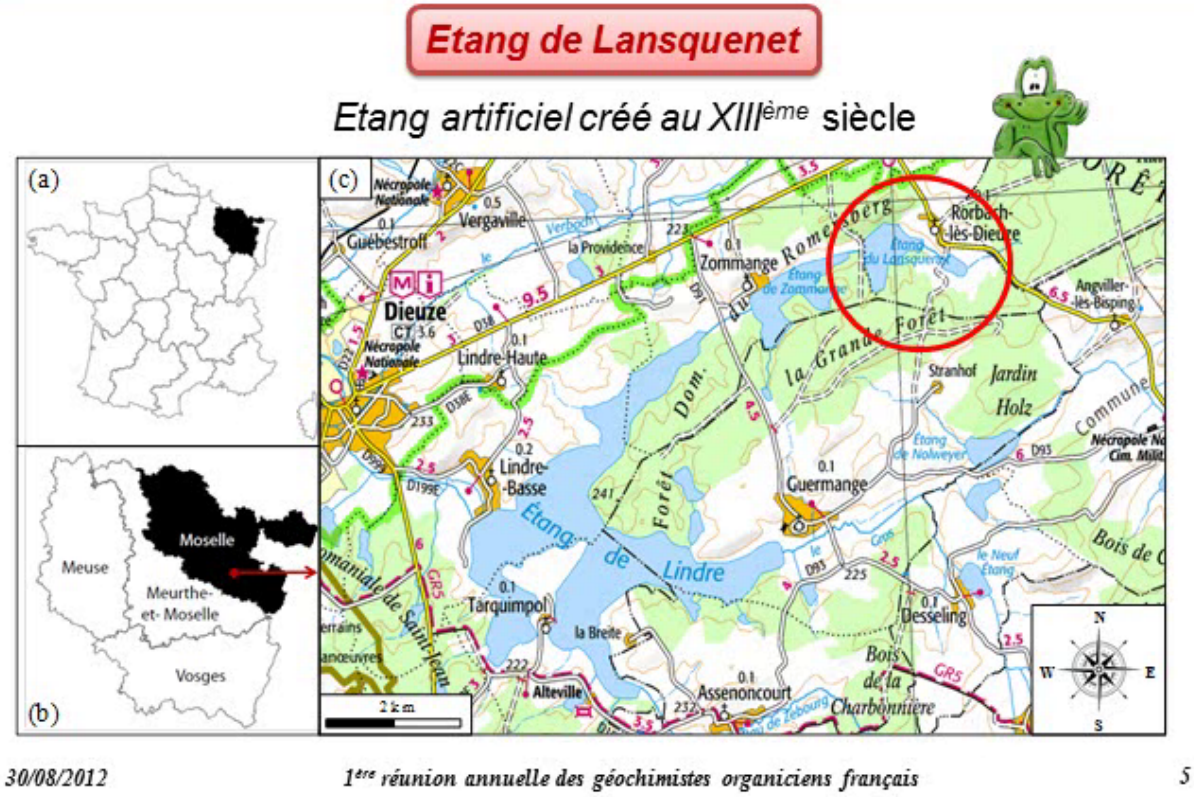


Figure 15 : *Diapositive avec bandeau (haut de la diapositive)*

L'examen des exposés oraux du corpus et en particulier des 4 exposés comportant un tel bandeau confirme l'aménagement de IMRAD et la place non négligeable accordée à l'exposé du contexte géographique ou géologique. Le tableau 8 reproduit les intitulés des parties explicitement mentionnées par un bandeau.

FR-S-O-01[93]	FR-S-O-02	FR-S-O-07	FR-S-O-04
Introduction	Introduction	Introduction	Introduction
Échantillons	Présentation de la zone d'étude	Friches industrielles	
Partie expérimentale	Matériels et méthode	Caractérisation globale des MOD Caractérisation moléculaire des MOD	Approche et protocole
Résultats	Résultats et discussion		Signature des sources
Perspectives	Conclusions	Perspectives futures	Nouveaux indices

Tableau 8 : *Macrostructure des 4 exposés qui affichent un bandeau*

Le découpage des articles écrits se réduit à l'oral à 5 phases principales :

- Introduction
- Contexte géographique / cadre d'étude
- Matériel / méthodes

93 Explicitation du codage : les premières lettres codent la langue (ici, le français), puis la discipline (ici, sciences) puis le mode (ici, oral) et enfin le numéro désigne le texte au sein de ce sous-corpus.

- Résultats / Discussion
- Conclusions / perspectives

Le tableau 8 met en évidence des dénominations diverses pour ces 5 phases, ce qui rend l'architecture d'un exposé individuel assez peu prédictible. On pourrait donc s'attendre à ce que les orateurs prévoient un guidage conséquent de leur auditoire, soit en annonçant le plan qu'ils vont suivre en début d'exposé, soit en ajoutant un bandeau à leur diapositive, soit en verbalisant la nature de ce dont ils vont parler, comme en (76), où l'orateur précise qu'il va évoquer les résultats de son étude :

(76) voilà donc j'en viens aux résultats euh [EIIDA - géochimie - oral]

Globalement, les exposés oraux n'annoncent pas majoritairement leur plan et n'ajoutent que parfois un bandeau. Mais les articles écrits ne sont pas plus obligeants envers leur lecteur et se dispensent de toute annonce de plan dans la même proportion que les exposés oraux, ce que montre la ligne *Rien* du tableau 9 : 8/15 soit 54 % pour les présentations orales, 9/15, soit 60 % pour les articles écrits, ne fournissent aucune structure explicite à leur destinataire.

	Présentations orales	**Article écrit**
Annonce plan Avec diapositive Sans visuel	 2 1	6
Bandeau	4	X
Rien	8	9
Total	15	15

Tableau 9 : *Guidage formel du lecteur à l'oral et à l'écrit*

Si la macrostructure de la présentation n'est que peu annoncée à l'avance et peu affichée sur les diapositives, l'orateur verbalise-t-il les « rubriques » de l'exposé, comme en (76), ci-dessus ? À vrai dire, pas davantage. Le seul moment charnière de l'exposé à être explicitement signalé de façon régulière est la conclusion. Sur les 15 présentations, 11 verbalisent le fait que le temps de l'exposé qui va suivre est la conclusion.

(77) voilà et donc pour conclure sur tout ce que je vous ai montré sur cette étude intégrée [EIIDA - géochimie - oral]

(78) donc pour conclure on a observé qu'on avait qu'on avait un effet ascendant de la composition de la matière organique des sédiments euh [EIIDA - géochimie – oral]

(79) donc euh en conclusion en fait la ... fluorescence c'est un outil qui est très intéressant au niveau spéléothème pour suivre la matière organique qui provient des sols [EIIDA - géochimie – oral]

Pendant tout le déroulement qui a précédé, une caractéristique de ces présentations orales est de dérouler le fil de l'étude de façon relativement chronologique, sans procéder de façon aussi claire et visible que dans les articles écrits à la reconstruction « après-coup » des temps logiques et non chronologiques de la recherche.

Pour vérifier s'il s'agissait là d'un effet de la discipline ou d'un effet du médium, je me suis intéressée aux présentations orales en linguistique, pour lesquelles nous ne disposons

malheureusement pas des diapositives projetées, lorsqu'il y en a, ainsi qu'aux articles écrits de linguistique. Le guidage du destinataire dans la macrostructure y est encore plus faible : 5 articles et 2 présentations orales seulement annoncent le plan de l'exposé. À l'écrit, il est lisible à travers les intertitres, mais à l'oral, l'auditeur chemine avec l'orateur sans possibilité d'anticipation.

Cette étude met en évidence le pouvoir structurant considérable des intertitres, qui n'ont finalement pas de contrepartie évidente à l'oral. Des moyens visuels notamment existent, tels qu'un bandeau ajouté aux diapositives, mais sont peu employés. En fait, c'est la logique même de l'exposé, qui semble à l'oral modifiée : plus linéaire, plus narratif, plus interactif.

4.5 Synthèse-bilan du chapitre

J'ai dans ce chapitre tenté de retracer le cheminement de plusieurs strates successives pour l'étude d'objets textuels propres à l'écrit, que nous avons d'abord désignés comme « titres de section » avant de les dénommer « intertitres », afin de bien les distinguer des titres qui précèdent et couvrent la totalité d'un texte, tels que les titres de livres ou d'articles. La nouveauté de l'objet d'étude a impliqué de penser un cadre d'analyse, donc de sélectionner les références théoriques les plus à même de nous permettre de penser les caractéristiques des intertitres. Les travaux en psychologie cognitive mettent en évidence le rôle positif des intertitres sur la compréhension des textes mais n'analysent pas finement leurs fonctions. Les théories du discours n'intègrent pas à leurs analyses la dimension visuelle des textes. Le modèle d'architecture textuelle permet de rendre compte de cette dimension visuelle mais reste lui aussi très restreint pour la compréhension des fonctions des intertitres. Nous avons donc développé un appareillage théorique et pratique qui permet de cerner ces fonctions :

- définitions d'observables les plus formels possibles, pour limiter la subjectivité inhérente aux jugements sur les données (cf. 2.2.2) ;

- constitution de corpus cohérent avec les hypothèses de fonctions reliées aux genres textuels ;

- codage des observables pour chaque intertitre ;

- élaboration d'hypothèses de fonctions ;

- analyse statistique pour traduire les fonctions proposées en corrélats formels, i.e. des configurations de traits.

Les fonctions observées sur le corpus original ont été vérifiées sur d'autres corpus[94], ce qui a permis de poursuivre les analyses dans deux directions : la compréhension des opérations de mise en relief réalisées « localement » par les intertitres, leur participation à la rhétorique du texte scientifique, leur rôle spécifique de l'*écrit*, sans contrepartie à l'oral.

Dans la seconde partie qui suit, je reviendrai sur deux aspects de ce que j'ai exposé là : le caractère formel des observables et les spécificités de la langue écrite, qui sont tous deux liés à des objectifs applicatifs.

94 Et par d'autres que moi-même, voir ce mémoire de Master 2 Recherche réalisé sous ma direction : Assani, N. (2010). *Aspects structurels de documents médicaux : Rapport entre titres, sous-titres et textes*. Université de Strasbourg.

Partie II
...orientées vers les applications

Préambule : linguistique appliquée / impliquée

J'ai dans la partie I fait la synthèse des travaux que j'ai menés dans l'optique de la description linguistique. Je vais maintenant me tourner vers la dimension applicative de ces recherches. Et pour donner tout son sens à cette dimension applicative, ce préambule apportera quelques éclairages sur la question des applications de la science.

D'après (Habert, 2004), la dichotomie, redoublée d'une hiérarchisation, entre science dite fondamentale et science appliquée, entre le savant qui produit de la connaissance et l'ingénieur qui en déduit « les applications industrielles dont elles se révèlent porteuses » (Lecourt, 2015 : 22), a son origine dans le positivisme d'Auguste Comte et a des prolongements jusque dans le paysage scientifique actuel.

> Cette thèse a longtemps fait obstacle à une authentique pensée de la technique en France. La science des ingénieurs n'étant pas tenue pour une science « proprement dite » par le positivisme dominant, leur savoir ne serait qu'un savoir-faire qui se limite à organiser la coopération entre la théorie et la pratique. (Lecourt, 2015 : 22)

Dans ce qui ressemble fort à un plaidoyer pour l'outillage des linguistes, Habert propose de dépasser cette opposition et d'« articuler autrement théorie et technique » :

> Outiller la linguistique et les linguistes, c'est faire l'hypothèse d'une articulation et d'une dialectique différentes entre applications et modélisations (Habert, 2004, § 28)

Il s'agit de non plus penser une chronologie de la science qui serait 'd'abord des avancées en recherche fondamentale puis des applications déduites', mais « *incorporer les conditions d'application dans la théorie* » (Condamines & Narcy-Combes, 2015 : 13) et, en particulier pour B. Habert et son « linguiste à l'instrument », être en mesure, en tant que chercheur-e, de « mettre les mains dans le cambouis »[95] et penser des dispositifs techniques qui soient des dispositifs d'investigation tout autant que des dispositifs d'application.

La notion d'application me semble alors pouvoir être appréhendée selon deux points de vue complémentaires. D'une part, elle concerne cette opposition, dénoncée par Habert (2004, 2005), entre un chercheur qui s'inscrirait dans une stricte visée de connaissances sans réflexion sur leur « utilité » ou leur « applicabilité » et un ingénieur qui ne produirait que de la mise en œuvre technique, sans que celle-ci interroge aucunement ou fasse évoluer le socle de savoirs la supportant. D'autre part, elle questionne le chercheur dans son rapport à la cité et à la société dans son ensemble. Car les applications, lorsque application il y a, sont les retombées concrètes, dans les univers professionnels et/ou, en particulier pour les SHS, sociaux, des acquis d'une discipline. Sur le plan institutionnel, l'effort accepté par la société de financement des recherches doit être récompensé par des progrès mesurables socialement ; du point de vue du chercheur, on peut considérer que

> [l]e chercheur a un rôle à jouer dans la cité et la spécificité de la science reste dans la distanciation et les comptes à rendre à la communauté en termes communément explicitables et reconnus en prenant en compte les phénomènes culturels, subjectifs et irrationnels. (Condamines & Narcy-Combes, 2015 : 6)

Dit autrement, même s'il reste impératif de défendre l'idée de recherches qui ne soient pas en prise directe avec une utilisation, une application immédiates, une discipline ne peut s'exonérer

95 Je traduis par cette expression imagée et familière le concept de « dirty hands » que Habert (2004, § 11) emprunte à d'autres.

de faire, de façons diverses, la démonstration de la valeur ajoutée par les savoirs qu'elle produit à l'ensemble des connaissances d'une société et d'une époque données.

Les questions de la place dans la cité et de l'utilité sociale se posent en des termes assez proches pour une autre science humaine, la sociologie, souvent sommée de se justifier en tant que discipline en répondant à la question « à quoi sert la sociologie ? » :

> tout sociologue qui prétend faire œuvre scientifique et, par conséquent, défendre son indépendance d'esprit contre toute imposition extérieure à la logique de son métier, est amené un jour ou l'autre à défendre, discrètement ou rageusement, sa liberté à l'égard de toute espèce de demande sociale (politique, religieuse, économique, bureaucratique…). […] Contre les injonctions multiformes de production d'un « savoir utile », les savants ont toujours eu à lutter pour la « curiosité gratuite » ou la « recherche de la vérité » en elle-même et pour elle-même. (Lahire, 2004)

Le chercheur doit ainsi habilement naviguer entre sa liberté de chercheur, liberté de définir ses objets de recherche selon la logique propre à sa discipline et liberté de gérer la publication de ses résultats, et la légitime demande de la société qui le finance, pour esquiver les écueils de la docilité, voire de la servilité, et d'un applicationnisme non réfléchi.

La linguistique, avec la psychologie ou la sociologie, fait partie des sciences humaines qui sont de plus en plus sollicitées pour produire de la connaissance, des conseils, voire même des « audits » dans des sphères diverses, ce dont témoignent (Léglise *et al.*, 2006) puis, plus récemment, (Condamines & Narcy-Combes, 2015). Au champ séculaire de la lexicographie et de la description grammairienne se sont ajoutés de nombreux domaines dans lesquels l'expertise du linguiste est sollicitée, à travers des « commandes » émanant d'entreprises ou d'institutions diverses :

- études de terminologie et d'aménagement linguistique dans des domaines spécialisés ;
- observation de la perméabilité entre langue spécialisée et langue générale ;
- analyse des discours écrits dans des sphères professionnelles[96] ;
- analyse et détection automatique des discours extrêmes sur la Toile ;
- enseignement des langues, étrangères et maternelle.

La liste n'est bien sûr pas exhaustive, loin, très loin de là, mais même ces quelques champs indiquent, par leur diversité, que « la demande sociale, simple volonté de participation curieuse au savoir ou d'implication à la gestion de faits de société, émerge et requiert considération » (Léglise *et al.*, 2006).

Les réponses à ces demandes, aussi bien que les recherches qui ne sont pas guidées par des objectifs applicatifs, conduisent les linguistes à mettre en œuvre des techniques appropriées à la collecte, au traitement et à l'analyse des données. Ce n'est pas dans son fauteuil, dans sa bibliothèque que l'on explore les rouages du « parler ordinaire », pour reprendre un titre de Labov (1993) c'est « sur le terrain », en créant ou en détournant les outils qui vont rendre cette exploration possible et féconde.

J'ai dans la section 3.1 donné un aperçu de certains de ces outils, les Bases de Connaissances Terminologiques (Aussenac-Gilles & Condamines, 2000), qui offrent une vue sur les triades terme-concept-(con)texte et permettent de répondre à une demande d'ordre termino-ontologique. L'élaboration de tels outils repose sur un génie linguistico-informatique, comme il

96 Voir le projet ANR *Écritures* http://www.univ-paris3.fr/description-scientifique-anr-ecritures-96466.kjsp?RH=1295620557102 page consultée le 14 aout 2017.

y a du génie chimique ou du génie mécanique. Non réflexion techniciste mais recherche de la meilleure façon de modéliser les connaissances recélées par les textes. L'outil, devenu instrument (Habert, 2005), est alors ce qui permet de voir, d'organiser, et les données, et l'analyse ou les analyses que le chercheur en produit.

De même que le microscope a permis l'observation de faits nouveaux, les outils d'exploration textuelle tels que les concordanciers livrent des rapprochements, des continuités, des distances que le linguiste ne peut apercevoir autrement (Pincemin, 2007). La mise au point de tels outils est partie intégrante de la recherche linguistique et répond à mon sens tout à fait à l'exigence de déplacement du rapport entre science et technique que Habert (2005) appelle de ses vœux et qui ne consiste pas strictement en une *application* ni de l'informatique, ni de la linguistique (Heiden, Magué & Pincemin, 2010). Ces outils reposent sur des prétraitements des textes, segmentation, lemmatisation, étiquetage, dont la réussite réclame à la fois une conceptualisation linguistique et une modélisation informatique abouties. En ce sens, il est bien plus pertinent d'envisager les réalisations concrètes qui peuvent découler des descriptions et des théorisations comme les fruits d'une collaboration de recherche que dans une division des rôles *chercheur / ingénieur*.

Mais une telle collaboration suppose une connaissance des possibles et une prise en compte des besoins de chacune des disciplines contribuant à la réalisation concrète (qu'elle soit un outil, un dispositif d'analyse, un dispositif de recueil de données, un dispositif d'enseignement…). Les descriptions linguistiques doivent être formulées en envisageant leur prolongement vers une application, ce qui suppose une compréhension et une anticipation de ce qui sera nécessaire pour l'application projetée – ou juste imaginée.

Dans les deux chapitres qui vont suivre, j'illustrerai de façons différentes cette articulation entre description et applications. Mon parcours m'a conduite à explorer successivement les deux orientations historiques de la linguistique dite « appliquée » (Léon, 2015) : le traitement automatique et l'enseignement de la langue.

Chapitre 5 - Vers le traitement automatique

J'ai mentionné en filigrane au cours des chapitres précédents l'optique de traitements automatiques dans laquelle certaines de mes (nos) descriptions ont été formulées. En particulier, en ce qui concerne les travaux sur les intertitres, le choix de privilégier des traits formels était lié à l'hypothèse d'une future implémentation d'une catégorisation automatique des intertitres. Les traits formels sont plus « calculables » par un système automatique qu'une catégorisation globale reposant sur l'interprétation. Cette catégorisation automatique n'a jamais vu le jour, faute de temps pour la mettre au point et faute d'un projet plus global dans lequel elle aurait trouvé son sens. Mais la démarche était typiquement l'approche du traitement automatique que je souhaite défendre ici : un travail de description appuyé sur corpus, la traduction des éléments jugés pertinents en critères calculables à peu de frais par un programme informatique ou un logiciel, l'évaluation des résultats, non nécessairement à l'aune d'un « gold standard »[97] mais en fonction de ce pourquoi le traitement est mis au point.

Cette démarche a été adoptée dans trois contextes différents, qui constitueront la matière des trois sections de ce chapitre : l'analyse syntaxique automatique, les relations lexicales, les ressources pour l'aide à la rédaction.

5.1 Analyse syntaxique automatique : Easy et Syntex

J'évoquais en préambule les prétraitements à appliquer aux textes d'un corpus pour rendre ensuite possible son exploration par un outil dédié tel que par exemple TXM (Heiden *et al.*, 2010). Typiquement, l'un des traitements les plus répandus consiste à segmenter le texte d'abord en phrases puis en « tokens » et à étiqueter chaque *token* avec des informations morphosyntaxiques. La figure 16 montre un exemple d'un tel étiquetage réalisé par MElt (Denis & Sagot, 2012) sur la phrase « Le théâtre met en évidence la réalité et la vérité », la sortie de l'étiqueteur a été reformatée en xml par un script *ad-hoc*. Chaque unité est encadrée par une balise « w » dans laquelle sont précisés la catégorie grammaticale (POS = *part of speech*) assignée par l'étiqueteur et le lemme (*lemma*) de l'unité, c'est-à-dire la forme canonique, dégagée de toute flexion, qui représente l'unité.

```
<sent>
  <w pos="DET" lemma="le">Le</w>
  <w pos="NC" lemma="théâtre">théâtre</w>
  <w pos="V" lemma="mettre">met</w>
  <w pos="P" lemma="en">en</w>
  <w pos="NC" lemma="évidence">évidence</w>
  <w pos="DET" lemma="le">la</w>
  <w pos="NC" lemma="réalité">réalité</w>
  <w pos="CC" lemma="et">et</w>
  <w pos="DET" lemma="le">la</w>
  <w pos="NC" lemma="vérité">vérité</w>
  <w pos="PONCT" lemma=".">.</w>
</sent>
```

Figure 16 : *Exemple d'étiquetage morphosyntaxique*

97 Un « gold-standard » est une référence à laquelle comparer les résultats des systèmes.

Un niveau d'annotation supérieur est atteint lorsque, en plus de cette information associée à chaque unité, sont annotées les relations syntaxiques entre les unités, telles que ici :

- *théâtre* a pour déterminant *le* ;

- *le théâtre* est le sujet du verbe *mettre* ;

- le verbe *mettre* constitue un groupe avec les constituants qui le suivent, *en évidence* ;

- *la réalité et la vérité* forme un groupe relié par coordination ;

- ce groupe relié par coordination est objet du groupe verbal *met en évidence* ;

- etc.

On voit avec ce simple exemple surgir de redoutables questions et décisions, par exemple : comment traiter *met en évidence*, comme une seule unité – ce qui suppose de l'étiqueter en bloc –, comme un syntagme – ce qui suppose d'étiqueter chaque élément et d'annoter ensuite des relations entre ces éléments ? Comment représenter la coordination ? Qu'est-ce qui est l'objet de *met en évidence*, chacun des deux SN *la réalité la vérité* ou un groupe de plus haut niveau construit par la coordination ?

Chaque analyseur syntaxique résout ces questions, et bien d'autres, d'une façon qui lui est propre, qui tient au cadre théorique auquel se réfère le concepteur de l'analyseur et aux utilisations qui seront ensuite faites de l'analyse produite. Comment alors juger de la performance comparée de divers analyseurs ? C'est un peu comme si l'on voulait évaluer des cuisiniers qui l'un propose une cuisine française, l'autre une cuisine thaï, l'autre une cuisine iranienne… Pour savoir lequel est « le meilleur », une bonne manière de faire est de donner à chacun les mêmes ingrédients et de leur demander de réaliser les mêmes plats. C'est un peu, *mutatis mutandis*, la démarche de la campagne EASy[98] (Évaluation des Analyseurs Syntaxiques) : confronter plusieurs analyseurs sur les mêmes données, assorties d'un même jeu d'étiquettes, en vue de produire l'analyse des mêmes relations syntaxiques – une part non négligeable du projet a ainsi consisté à définir les relations syntaxiques visées, au nombre de 14 (Paroubek *et al.*, 2005). Ma participation à ce projet a été une occasion de contribuer au développement de Syntex (Bourigault & Fabre, 2001 ; Bourigault *et al.*, 2005) qui était alors en constante évolution à l'ERSS [10,11].

L'esprit de Syntex était loin des traitements statistiques et des démarches d'apprentissage désormais communes en TAL (traitement automatique des langues), dont l'expansion fait écrire à J. Léon :

> Dans la période actuelle, la mise à distance du TAL à l'égard de la LA [linguistique appliquée] participe d'un mouvement plus général d'éloignement du TAL à l'égard de la linguistique. La prédominance des méthodes statistiques en TAL enlèvent leurs prérogatives aux domaines de la linguistique traditionnellement sollicités pour l'élaboration de règles (syntaxe, morphologie ou sémantique). La linguistique est au mieux instrumentalisée et les structures linguistiques quand on les utilise encore sont réduites à de simples données, au même titre que les nombres, les tableaux, et les « sacs de mots ». Le TAL ne s'intéresse plus à la linguistique appliquée, parce qu'il ne se compte plus parmi les applications de la linguistique. (Léon, 2015, § 36)

Dix ans plus tôt, Syntex était un analyseur fondé sur des connaissances linguistiques et résolvant les problèmes posés à l'analyse syntaxique par des règles destinées à capter la

98 http://www.technolangue.net/article.php3?id_article=198 page consultée le 14 aout 2017.

subtilité des variations des relations de dépendances selon les contextes. Un exemple très trivial de ces variations concerne les prépositions et leur faculté de se rattacher à des groupes divers, parfois très éloignés, ce qui complique singulièrement l'analyse automatique[99]. Considérons (80) et (81), qui sont des transformations libres d'une phrase extraite du journal Le Monde.

(80) Le chancelier assure l'Assemblée nationale de la continuité de sa politique franco-allemande. [exemple fabriqué]

(81) Le chancelier assure le maintien indéfectible de la continuité de sa politique franco-allemande. [exemple fabriqué]

Dans (80), *de la continuité de sa politique franco-allemande* se rattache au verbe, selon le schéma syntaxique « assurer quelqu'un de quelque chose » ; dans (81), le même constituant se rattache au nom qui précède *maintien*, selon le schéma « assurer (= garantir) quelque chose », comme dans (82).

(82) À chaque période, l'épargne agrégée assure le financement du stock de capital de la période suivante. [Termith]

Une solution pour de telles ambigüités est de s'appuyer sur un apprentissage endogène, c'est-à-dire une première analyse du corpus qui construit automatiquement les schémas valides à partir des contextes dans lesquels aucune ambigüité n'est observée (Frérot, Bourigault & Fabre, 2003), dont (83) donne un exemple, apprentissage complété par un lexique syntaxique qui fournit de tels schémas (Frérot, 2005).

(83) L'Assemblée Nationale est assurée de la continuité politique. [exemple fabriqué]

Un problème analogue se pose pour *que*, je l'ai traité dans (Jacques, 2005b) [11]. Selon la catégorie grammaticale de *que*, la relation de dépendance ne sera pas la même, y compris parfois dans des contextes qui se ressemblent.

(84) la difficulté première est de faire comprendre aux acteurs **que** les problématiques induites par la réforme de la PAC non seulement concernent l'ensemble du « monde agricole » [Termith]

(85) Les types d'acteurs **que** j'ai utilisés dans ces exemples (individus, organisations, etc.) correspondent plus ou moins à des niveaux de masse [Termith]

En (84) comme en (85), *que* suit le nom *acteurs*, cependant, dans le premier cas, *que* se rattache au verbe *comprendre* en tant que conjonction de subordination tandis que, dans le second cas, il faut le rattacher au nom en tant que pronom relatif.

Pour traiter convenablement *que*, il faut différencier ces deux cas de ses autres utilisations :
- adverbe dans une négation exceptive (Gaatone, 1999) :

(86) On **ne** trouve **que** très peu d'exemples de ce type [Termith]

- conjonction dans un système comparatif ou corrélatif (Riegel, Pellat & Rioul, 2009 : 868) :

(87) les enjeux sont **autant** culturels **que** sociaux [Termith]

99 Je ne résiste pas au plaisir de signaler qu'il n'y a pas que l'analyse automatique à être mise en défaut pour l'interprétation des groupes prépositionnels, considérer par exemple l'étiquette « *Crème de marrons avec morceaux du Massif central* » qui laisse quelque peu perplexe – à consulter ici http://bling.hypotheses.org/2203, avec quelques autres dont on aura un aperçu grâce aux savoureux billets de BLING (Blog de Linguistique Illustrée) consacrés à ce sujet : http://bling.hypotheses.org/1948

(88) E. Bautier et J.-Y. Rochex (2004) évoquent le contrôle de l'activité des élèves en difficulté grâce à des tâches **tellement** simplifiées **qu'**elles ne font appel qu'à des traitements de bas niveau. [Termith]

L'aperçu des fonctions de *que* est ici assez sommaire quand on sait que la définition lexicographique de *que* dans le TLFi couvre l'équivalent de plus de quinze pages au format A4, inventorie les trois catégories grammaticales indiquées (conjonction, pronom, adverbe) et liste pour chacune quantités de divisions et de subdivisions afin d'en cerner tous les emplois. Cette unité est donc parmi les plus polycatégorielles et polyfonctionnelles de la langue.

Une analyse syntaxique correcte de *que* repose alors crucialement sur l'identification de sa catégorie grammaticale, c'est celle-ci qui dicte de chercher ensuite dans la phrase l'élément auquel le rattacher : un nom, un verbe… Or une comparaison de l'étiquetage produit par Treetagger[100] pour *que* avec un étiquetage de référence vérifié manuellement – du moins est-ce ce qu'affirme la documentation du corpus utilisé, le corpus CRATER[101] – montre un taux d'erreur conséquent : pour les 1183 *que* relevés dans CRATER, 75 % sont correctement étiquetés, ce qui implique que 25 %, soit 1 sur 4, ne reçoivent pas la bonne étiquette (Jacques, 2005b : 140) [11]. Il fallait mettre au point une stratégie pour à la fois améliorer l'étiquetage de l'unité dans son contexte, en modifiant si nécessaire la catégorie proposée par l'étiqueteur (Treetagger), et proposer une analyse syntaxique convenable, c'est-à-dire qui satisfasse les attentes établies pour EASy : rattacher l'adverbe ou la conjonction au bon verbe ou au bon nom en cas de complétive, rattacher le pronom relatif au bon nom ou au bon pronom.

L'enjeu ici est donc de déterminer à quels indices formels se fier pour prendre une décision sur la catégorie et sur l'unité avec laquelle établir une relation de dépendance pour *que*. C'est un problème intéressant pas seulement du point de vue du traitement automatique mais aussi du point de vue linguistique car il met en évidence l'endroit où seule la compréhension humaine permet de trancher[102]. Mais pour les traitements automatiques, il faut se passer de compréhension et définir ce qui pourrait pallier cette absence. Une observation des contextes, dans l'optique d'inventorier les traits des différentes constructions possibles pour *que* et d'y repérer les indices formels pertinents, m'a permis d'élaborer une stratégie à base de règles reposant sur la prise en compte de la partie de la phrase qui précède l'occurrence de *que*.

Pour comprendre cette stratégie, il faut savoir que l'analyseur fonctionne de façon modulaire : chaque phrase est traitée plusieurs fois, par des modules qui chacun peuvent s'appuyer pour leur part du travail sur le résultat de l'analyse effectuée par les modules précédents. Au moment

100 L'étiqueteur Treetagger est l'un des plus utilisés pour l'étiquetage morphosyntaxique automatique, il est disponible gracieusement et dispose de modules d'entrainement permettant son adaptation à de nouvelles langues et de nouveaux corpus. Voir http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ page consultée le 16 aout 2017.

101 Le corpus CRATER aligne des textes du domaine des télécommunications dans trois langues, français, anglais, espagnol. Il est disponible via l'agence ELRA, qui le décrit dans ces termes : « L'offre consiste en 3 fois 1 million de mots en anglais, français et espagnol avec des annotations morphosyntaxiques (vérifiées par un opérateur humain). » http://catalog.elra.info/product_info.php?products_id=84&language=fr page consultée le 16 aout 2017.

102 Il me semble qu'il y aurait là matière à rapprochement avec ce que je traiterai dans le chapitre suivant : l'enseignement de la langue. Je fais ici l'hypothèse que les cas qu'un programme informatique résout uniquement sur la base d'un calcul formel sont ceux qui poseront le moins de difficultés à un locuteur allophone en situation d'apprentissage tandis que les cas que le programme ne résout pas sont ceux qui manifestent une forme d'ambigüité syntaxique et poseront donc davantage de problèmes à un allophone. Mais c'est en l'état une hypothèse très embryonnaire qui mériterait plus d'attention pour estimer si elle vaut d'être travaillée plus avant.

où intervient le module de traitement de *que*, la plupart des relations de dépendance de la phrase ont déjà été déterminées et le module peut les exploiter.

L'analyse repose sur deux étapes : dans la première, le module cherche des indices des structures inventoriées au préalable. La nature ou la forme des unités qui précèdent *que* servent de déclencheur pour la mise en œuvre de telle ou telle règle. Par exemple, si le *token* avant *que* est un nom ou un adjectif, il faudra déterminer si *que* est un pronom relatif ou une conjonction par l'application d'une certaine règle, si parmi les *tokens* qui précèdent *que* on trouve *même* ou *autant* ou un autre marqueur potentiel de structure comparative, une autre règle s'appliquera. Cette première étape, qui explore le contexte dans la proximité immédiate de *que*, permet de découvrir essentiellement des dépendances très proches : adverbe ou conjonction introduisant une complétive immédiatement précédés du verbe (89), structure comparative aux éléments adjacents (90) :

(89) Les linguistes **admettent que** les structures sémantiques (qu'elles soient riches ou minimales) sont sous- spécifiées du point de la référence [Termith]

(90) le public n'est pas le **même que** celui d'un musée ou d'un autre établissement à vocation touristique [Termith]

Est opérée ici une *déduction positive* (Vergne & Giguet, 1998) : le contexte permet de déduire que *que* est bien une conjonction ou un adverbe. Cette première étape améliore l'étiquetage de 14 points : on passe de 75 % de *que* correctement étiquetés à 89 %.

La deuxième étape va plus loin que le contexte immédiat et exploite la totalité de l'analyse de la phrase pour vérifier des hypothèses de structures. Par exemple, il s'agit de découvrir en (91) une structure comparative dont le marqueur (l'adjectif *mêmes*) ne précède pas immédiatement *que* mais est intégré à un syntagme nominal.

(91) les entreprises nouvellement créées ne sont pas tenues aux **mêmes** obligations de déclarations fiscales **que** les autres entreprises [Termith]

De même, en (92), la découverte de la structure corrélative nécessite de « sauter par-dessus » le SN qui précède *que* pour repérer le marqueur de cette structure (l'adverbe *autant*).

(92) ce baptême représente **autant** un rite d'institution **qu'**un rite de passage [Termith]

Si aucun indice n'est concluant, l'étiquette posée par Treetagger lors de l'étiquetage initial est laissée en l'état : l'objectif était d'exploiter l'analyse syntaxique pour augmenter la qualité de l'étiquetage, si elle ne permet pas de remettre en cause les décisions prises par le *tagger*, il est préférable de s'abstenir de toute modification. Cette seconde étape améliore moins spectaculairement l'étiquetage puisqu'elle permet de gagner seulement 3 points en portant le nombre de *que* correctement étiquetés à 92 %.

Dans le même temps, le module doit procéder à l'annotation des relations de dépendance : *que* adverbe de négation exceptive est sous la dépendance d'un verbe (une seule dépendance à annoter), *que* conjonction de subordination introduisant une complétive est sous la dépendance d'un verbe ou d'un nom et a sous sa dépendance le verbe de la proposition qu'il introduit (deux dépendances à annoter), etc.

Au final, si l'on prend en considération uniquement les *que* pour lesquels l'étiquetage est correct[103], c'est un total de 1843 relations qui doivent être annotées, tous types confondus.

103 Cette restriction répond au souci d'évaluer la performance des règles mises au point pour l'annotation, pas la performance globale du programme. On ne peut demander à un module de calculer des relations

Deux mesures de performance permettent d'apprécier la réussite de l'annotation : le rappel, qui concerne le nombre de relations effectivement étiquetées sur le nombre de relations que l'on devait étiqueter, la précision, qui concerne le nombre de relations correctement étiquetées sur le nombre de relations étiquetées. Le module présente des résultats tout à fait honorables puisqu'il obtient un rappel de 93% (ce qui signifie que dans 7 % des cas, il aurait dû annoter une relation et il l'a manquée) et une précision de 94 % (ce qui signifie que dans 6 % des cas, la relation qu'il a annotée ne convient pas). Les 6-7 % de relations manquantes ou erronées peuvent paraitre une réussite insuffisante à un linguiste, mais si l'on considère les performances habituelles des analyseurs syntaxiques, on constate que ces taux sont acceptables et même satisfaisants. Dans sa thèse qui décrit l'élaboration d'un analyseur combinant méthodes statistiques et méthodes symboliques (c'est-à-dire règles), Urieli (2013 : 145) rapporte un taux de précision initial de 87,27 %, toutes relations confondues – et note l'analyse spécifique de *que* comme un vecteur d'amélioration de l'analyse globale (voir aussi Urieli, 2015)[104].

Les règles appliquées dans ces deux étapes ont été construites à partir d'une forme de modélisation des diverses constructions, modélisations qui impliquent de déceler des structures communes sous d'éventuelles variations de surface et de savoir de quels éléments il convient de faire abstraction et sur quels éléments il faut au contraire fonder la méthode de reconnaissance et de traitement des « indices ». La linguistique est ici « appliquée » dans le sens où la description produite ne vaut pas pour elle-même – il ne s'agit pas au premier chef d'augmenter la connaissance du système de la langue – mais représente une des étapes de la tâche. Cette description orientée vers un but peut néanmoins permettre d'éclairer les fonctionnements linguistiques. C'est ce dont il va être question dans la section qui suit, qui évoque divers travaux autour de « patrons lexico-syntaxiques ».

5.2 Identification en corpus de relations lexicales par des patrons lexico-syntaxiques

J'ai évoqué dans le chapitre 3 les travaux menés en terminologie. Je vais revenir ici en particulier sur ceux qui traitent de l'extraction de connaissances à partir de textes et des moyens les plus appropriés de représentation de ces connaissances.

À côté des recherches sur les termes, de nombreuses études portent sur la structuration des termes et des concepts en un système cohérent, qui représente les relations pertinentes entre ces termes (cf. Grabar & Hamon, 2004). J'ai donné un aperçu de relations avec la figure 1, page 61, qui schématise le réseau de relations autour d'une même entité, *ASF*, pour trois acteurs de la gestion des déplacements. La relation qui s'établit entre *ASF* et d'autres termes du domaine est pour deux d'entre eux une relation d'hyperonymie. Il s'agit d'une relation structurante essentielle pour rendre compte de l'organisation de la conceptualisation d'un domaine, dans le sens où elle participe de la mise en ordre et de la catégorisation de notre domaine d'expérience, en ce qu'elle permet de construire des taxinomies. Pour le troisième, la relation est de *partie à tout* (méronymie, cf. Cruse, 1986), dont les divers types ont été finement décrits par (Winston, Chaffin & Herrmann, 1987). Ces deux relations sont largement

correctes si les informations dont il dispose pour ce faire sont fausses.

104 Le travail d'équipe pour cette campagne a été payant, Syntex s'est avéré le meilleur analyseur pour les genres textuels sur lesquels il a concouru – le choix a été fait de ne pas produire des analyses sur tous les genres textuels, en faisant l'hypothèse que les performances de l'analyseur seraient très dégradées sur les genres très éloignés de ceux ayant servi à sa mise au point. Cette question est précisément traitée dans la section 5.2.1.

exploitées dès lors qu'il s'agit de rendre compte de la façon dont les acteurs d'un domaine spécialisé le structurent.

De même que pour la terminologie, le paradigme dans lequel j'ai travaillé, avec N. Aussenac-Gilles (Aussenac-Gilles & Jacques, 2006, 2008 ; Jacques & Aussenac-Gilles, 2006) [16, 17, 19] et avec A. Condamines (Condamines & Jacques, 2006) [14], est textuel : il s'agissait de découvrir, dans les textes, les formulations de telles relations et de les modéliser pour en faire des outils de recherche utilisables au-delà du corpus d'étude. Ces deux recherches complémentaires ont eu l'intérêt de permettre chacune un retour sur le fonctionnement de la langue.

5.2.1 Marqueurs de relations : mener une réflexion sur les genres

Dans la section 3.1, j'ai mentionné les travaux autour des Bases de Connaissances Terminologiques, bases de données destinées à engranger les informations pertinentes pour enregistrer ce que les textes livrent de la connaissance d'un domaine : les termes, les concepts, les relations qui structurent l'ensemble. Un complément de ces BCT est Caméléon (Aussenac-Gilles & Séguéla, 2001), dont le développement est poursuivi par N. Aussenac-Gilles (Aussenac-Gilles & Jacques, 2008) [17, 19]. Caméléon propose dans sa version 3 d'associer en un même outil 1. un module de recherche en corpus des relations susceptibles d'aider à la modélisation d'un domaine et 2. un module de construction du modèle du domaine. Notre collaboration a porté sur le premier module, qui était au début une coquille vide et qu'il fallait « peupler » avec des marqueurs de relations.

La conception sous-jacente à cette notion de marqueurs de relation tient à l'hypothèse de l'existence d'un nombre fini d'expressions indiquant de façon régulière une certaine relation entre deux termes. L'association des éléments constituant l'expression *marque* la relation. Par exemple, les deux phrases suivantes commençant par « Parmi les » permettent de comprendre que, respectivement, Lewin est un pionnier (93) et que les Mandé sont des Nigritiens (94).

(93) **Parmi les** pionniers, Lewin dès 1959 a notamment montré qu'en réunissant des sujets et en favorisant les interactions entre ces sujets sur un thème donné, ces derniers étaient engagés individuellement […] [Termith]

(94) **Parmi les** Nigritiens, les Mandé « représentent la race nigritienne, avec, souvent un mélange plus ou moins fort de sang éthiopien ou berbère. [Termith]

La constitution d'une base de données regroupant toutes les expressions des diverses relations visées suppose en premier lieu de dresser une liste des expressions à rechercher, puis de mettre en évidence les régularités lexicales et syntaxiques pertinentes. J'ai cité dans le chapitre 2 plusieurs travaux de recherche menés dans cette perspective, pour rappel, (Borillo, 1996 ; Hearst, 1992) se sont intéressées à l'expression de la relation d'hyperonymie, (Jackiewicz, 1996) à la relation partie-tout.

L'objectif de Caméléon 3 est de fournir aux utilisateurs une base préétablie de marqueurs supposés génériques, c'est-à-dire indépendants du domaine à modéliser. On considère qu'il existe des relations spécifiques à un domaine qui sont à déterminer au cas par cas. Ces relations spécifiques à un domaine peuvent être par exemple une relation de causalité telle que celle qui s'établit en médecine entre un trouble et un symptôme, comme en (95).

(95) l'artère est exposée à des traumatismes causés par le contenu luminal, **ce qui provoque** une hémorragie [Médecine]

Dans la mesure où elles sont dépendantes du domaine, il revient à chaque utilisateur de Caméléon de les définir et les décrire pour son propre domaine. Mais, pour « amorcer » sa modélisation, il doit disposer du moyen de retrouver déjà dans les textes de son domaine les contextes dans lesquels s'expriment les relations génériques telles qu'hyperonymie et partie-tout. C'est là la destination de la base de marqueurs génériques : donner un ensemble de requêtes « prêtes à l'emploi », à priori applicables sur tout corpus, sous la forme de patrons lexico-syntaxiques, c'est-à-dire d'une combinaison d'éléments lexicaux et de contraintes syntaxiques. Par exemple, pour capter un énoncé tel que « Le canyon est une vallée à flancs raides », Rebeyrolle et Tanguy (2001 : 169) formulent le patron suivant

être (Adv)1 Det|Num 2 (Nom sauf L)

qui se traduit par 'le verbe être, suivi facultativement d'un adverbe, puis d'un déterminant ou un numéral, de deux mots quelconques facultatifs, et enfin d'un nom, à l'exclusion des noms d'une liste préétablie (*cas*, *exemple*, etc.)'. Un tel patron est supposé pouvoir capter une relation d'hyperonymie dans n'importe quel corpus.

C'est justement le caractère supposé de généricité qui est à questionner : est-il avéré ou n'est-il que la résultante d'une représentation partielle des fonctionnements langagiers ? En TAL comme dans d'autres domaines qui privilégient les programmes informatiques comme mode d'accès au texte – ce qui se justifie par les tâches visées : recherche d'informations, extraction d'informations, extraction de connaissances, et par la quantité de données à traiter –, le texte est souvent traité de façon indifférenciée comme un « réservoir » (de mots, d'informations, de connaissances…). Les variations selon le genre textuel sont restées peu étudiées jusqu'au début des années 2000 et il est frappant de voir que ce sont surtout les linguistes qui s'intéressent à la question, par exemple Malrieu et Rastier (2001) explorent de façon systématique la variation des catégories morphosyntaxiques en fonction du genre textuel.

Cela tient à mon sens au fait que les domaines qui traitent du texte pour les tâches évoquées plus haut mènent surtout des recherches sur les moyens à mettre en œuvre pour améliorer les performances, ce qui suppose de s'intéresser plus aux méthodes et à la façon d'organiser les tâches qu'aux données. Pourtant, il a été montré qu'une même tâche ne donne pas les mêmes résultats selon les données sur lesquelles elle est effectuée, (Illouz, 2000) montre un effet du type de texte sur les performances d'un étiqueteur, (Frérot, 2005) montre elle aussi une sensibilité au genre de texte des performances des stratégies de rattachement prépositionnel (voir section précédente, page 121). Nous nous sommes inscrites dans ce champ de recherches [16] en évaluant systématiquement les patrons lexico-syntaxiques destinés à repérer les relations conceptuelles dans huit corpus textuels délibérément variés, certains constitués pour les études présentées dans les chapitres précédents :

1. un guide de planification de réseau électrique (GDP, 187 800 mots) ;
2. des articles scientifiques de la conférence Ingénierie des Connaissances (IC, 198 500 mots) ;
3. des articles extraits de l'Encyclopedia Universalis, du domaine de la géomorphologie (ENC, 200 500 mots) ;
4. un manuel de géomorphologie (GEO, 260 000 mots) ;
5. un manuel de parapente (PAR, 23 800 mots) ;
6. un manuel de spécification de logiciels dans le domaine de l'électricité (MOU, 57 500 mots);
7. plusieurs thèses en archéologie (ARCH, 95 000 mots) ;
8. des textes du domaine de la télécommunication (CRATER, 817 000 mots).

70 patrons différents ont été appliqués à ces huit corpus : 18 sont des adaptations des patrons élaborés par (Rebeyrolle & Tanguy, 2001) pour le repérage des définitions, les autres sont des adaptations des patrons élaborés par (Séguéla, 1999), concepteur du premier Caméléon, et par N. Grabar : 35 pour la relation d'hyperonymie, 14 pour la méronymie, 1 pour les reformulations et 2 divers. Notre évaluation a porté sur deux mesures : la productivité du patron, c'est-à-dire le nombre de contextes qu'il permettait de retrouver, et sa précision, c'est-à-dire le nombre de contextes démontrant bien la relation recherchée. Nous n'avions pas de mesure de rappel, car il aurait fallu avoir une liste de tous les contextes exprimant des relations lexicales dans ces huit corpus, nous ne l'avions pas.

Le tableau 10, repris de (Jacques & Aussenac-Gilles, 2006 : 25) [16], donne un aperçu de ces deux mesures pour quelques patrons dans les 8 corpus (N est le nombre de contextes renvoyés par le patron, P est la précision du patron en pourcentage, les deux valeurs extrêmes de précision sont en gras, en excluant la précision calculée sur un nombre d'occurrences inférieur à 5).

Quelques exemples des contextes atteints aideront à mieux cerner l'objet traité, je mets en caractères gras les éléments des patrons qui permettent de reconnaitre ces contextes.

- *est-un*

(96) L'albédo d'un corps **est un** rapport qui exprime la partie de rayonnement directement réfléchie et donc non absorbée.

- *et Adv* (Adv = *notamment*, *notablement*, *spécialement*, *particulièrement*)

(97) En ce qui concerne les grandes stations **et particulièrement** les stations Intelsat de type A…

- *sorte de*

(98) les amines, qui sont des **sortes de** substances chimiques ;

- *inclure*

(99) Les services de base à fournir dans le Rmtp **comprennent** les téléservices et les services support…

- *partie de*

(100) l'ontologie est un **composant de** la mémoire d'entreprise…

- *situé dans*

(101) Le Ccm interroge l'Elv à chaque fois qu'il a besoin d'informations relatives à une station mobile donnée **située** à ce moment **dans** la zone du Ccm.

- *c'est-à-dire*

(102) la résolution, **c'est-à-dire** la taille des objets qui se distinguent, est de 100 m.

	GDP		IC		ENC		GEO		PAR		MOU		ARCH		CRAT	
	N	*P*[105]	N	*P*	N	*P*	N	*P*	N	*P*	N	*P*	N	*P*	N	*P*
être-un	268	*19*	574	*19*	420	*16*	752	*23*	62	***40***	129	***12***	181	*29*	751	[106]
et Adv	10	*10*	15	*7*	66	***5***	56	*30*	2	*0*	6	*17*	13	*38*	19	***58***
sorte de	0		7	***57***	1	*100*	3	*67*	0		0		0		4	***100***
inclure	75	*51*	32	*41*	29	***62***	16	*50*	2	*100*	18	*61*	27	***19***	267	*48*
partie de	0		0		1	*100*	7	***0***	1	*0*	0		1	*0*	11	***18***
situé dans	40	*53*	63	*38*	55	***24***	38	***24***	4	*75*	4	*50*	36	*56*	291	***59***
c'est-à-dire	6	*67*	37	*54*	14	***29***	40	***80***	2	*100*	3	*100*	8	*63*	11	*64*

Tableau 10 : *Résultats pour un échantillon de patrons*

105 Rappel : la précision est le rapport du nombre de contextes correspondant à ce que l'on souhaitait sur le nombre de contextes renvoyés par le patron. Une précision de 40 % signifie que 40 contextes sur 100 correspondent à ce que l'on voulait.

106 La précision n'a pu être calculée sur ce corpus : trop de contextes étaient incompréhensibles pour un non-spécialiste et nous n'avions pas de spécialiste des télécommunications dans notre entourage. C'est là une des limites du travail en corpus spécialisé, je reparlerai de l'analyse des textes par un non-spécialiste dans la section 5.2.2.

Comme le montre le tableau 10, les corpus sont de tailles différentes, il n'y a pas lieu de commenter la productivité (le nombre de contextes renvoyés par le logiciel) pour l'ensemble, mais on voit tout de même, sur les trois premiers corpus qui sont de taille comparable (de 187 000 à 200 000 mots), un rapport qui peut aller du simple au sextuple pour certains patrons. Le tableau 10 met surtout en évidence la disparité considérable sur la précision des patrons. Sans prendre en compte les quelques cas où le nombre de contextes obtenus est inférieur à 5, ce qui à mon sens donne une précision qui ne signifie pas grand-chose, on voit que le pourcentage de contextes pertinents varie lui aussi du simple au double, au triple, voire davantage : pour *être-un*, de 12 % à 40 %, pour *et Adv*, de 5 à 58 %.

Ces écarts conduisent à deux conclusions : 1. avec les mesures qui sont données, les traitements automatiques ne doivent pas négliger de décrire les caractéristiques des corpus sur lesquels ces mesures sont obtenues, 2. on connait finalement assez peu les paramètres des textes qui influent sur ces différences de mesures. On voit par exemple que les trois manuels (géomorphologie, parapente, spécification de logiciel) ne présentent pas des taux de précision similaires pour les divers patrons, loin de là, pas plus que les deux corpus de textes scientifiques, Ingénierie des Connaissances et Archéologie.

De telles disparités ont des conséquences du point de vue du traitement automatique des langues et du point de vue de la linguistique. En ce qui concerne le TAL, l'enjeu est la réplicabilité des traitements et de leurs performances. Quand un étiqueteur morpho-syntaxique annonce une précision moyenne de 96 %, l'utilisateur s'attend à ce que cette précision soit vérifiée, à un ou deux points près, sur son propre corpus. Si on indique que la précision du patron X (patron dit « générique », je le rappelle) est d'environ 40 % (par exemple *être-un* dans le manuel de parapente), on jugera très insatisfaisant de la voir chuter à 12 % sur un autre texte. Ce que notre expérimentation met en évidence, c'est cette variabilité de performances, assez peu acceptable dès lors que l'on n'est plus dans un contexte de recherche mais dans un contexte applicatif. Pour une société qui ferait de la veille documentaire ou terminologique, écarter 4 contextes non pertinents sur 10, ce n'est pas le même cout que d'en écarter 9 sur 10 (écart de précision pour le patron '*et Adverbe*'). Dans le premier cas, l'effort à consentir peut paraitre acceptable, dans le second, il est rédhibitoire.

Du point de vue de la linguistique, la question qui se pose est celle du rapport entre genres, types et fonctionnements textuels. Pour Branca-Rosoff,

> la notion de genre est une notion biface qui fait correspondre une face interne (les fonctionnements linguistiques) avec une face externe (les pratiques socialement signifiantes) (1999 : 116)

Et c'est bien là toute la difficulté, au genre comme pratique socialement située ne correspondent pas nécessairement et pas de façon déterministe, des fonctionnements linguistiques exclusifs, c'est-à-dire des fonctionnements linguistiques qui seraient spécifiques à un genre à l'exclusion d'un autre, de façon telle que l'on puisse poser une relation bi-univoque entre un genre avéré et certains traits linguistiques. Les traitements automatiques servent ici de révélateur : ils montrent que l'on ne peut prédire une identité de résultats pour une même tâche entre textes d'un même genre.

Une tentative de réponse à cette difficulté vient des typologies inductives qui prennent la question par « la face interne », pour reprendre l'expression de Branca-Rosoff ci-dessus, c'est-à-dire qui proposent des regroupements de textes à partir de leurs caractéristiques linguistiques. C'est tout le sens des travaux de Biber (1988, 1995) qui utilise des méthodes

d'analyse statistique pour regrouper les textes selon leurs traits linguistiques et qui interprète ensuite ces regroupements selon des critères situationnels :

> the term *text type* has been used in my own previous analyses to refer to text categories defined in strictly linguistic terms (Biber 1989). That is, regardless of purpose, topic, interactiveness, or any other non-linguistic factors, text types are defined such that the texts within each type are maximally similar with respect to their linguistic characteristics (lexical, morphological, and syntactic), while the types are maximally distinct with respect to their linguistic characteristics. After the text types are identified on formal grounds, they can be interpreted functionally in terms of the purposes, production circumstances, and other situational characteristics shared by the texts in each type. (Biber, 1995 : 10)

Pour autant, la mise en correspondance des types ainsi obtenus et des réponses des textes aux traitements automatiques qui leur sont appliqués reste à faire. Il n'est en outre pas sûr que des textes regroupés par leurs traits linguistiques et qui présenteraient des résultats similaires pour une tâche donnée, réagissent de façon tout aussi similaire pour une autre tâche. Il semblerait donc judicieux, lorsque l'on met au point un traitement automatique, d'apporter des indications sur les corpus testés telles qu'un utilisateur du traitement puisse apprécier la distance de son propre corpus avec ceux qui ont été expérimentés, mais une telle documentation requiert la définition des paramètres à faire entrer dans la description. La tendance actuelle dans le TAL français de mobiliser de plus en plus des méthodes statistiques sur de gros volumes de données supposés permettre de « lisser » l'hétérogénéité des textes l'éloigne de cette réflexion sur les types et sur les traits linguistiques pertinents.

Les approches symboliques semblent avoir fait leur temps pour les traitements automatiques et avec elles la place de la linguistique dans le TAL. Elles semblent toutefois pertinentes pour des tâches dans lesquelles le TAL est conçu comme une assistance, un outil destiné à faciliter une analyse effectuée en dernier ressort par un être humain, ce qui était le cas du projet autour de Caméléon 3.

La dernière étude en relation avec l'extraction de connaissances que je vais décrire, menée avec A. Condamines, met en évidence la part irréductible de l'intervention humaine pour traiter convenablement les textes.

5.2.2 Relation d'hyperonymie : retour au fonctionnement textuel

A. Condamines travaille de longue date sur la mise au point de marqueurs linguistiques de ces relations lexicales génériques telle qu'hyperonymie et méronymie (voir par exemple Condamines, 2000, 2006). Une utilisation de marqueurs telle que je viens de la décrire, c'est-à-dire par projection de patron lexico-syntaxique ou encore par recherche d'une unité lexicale telle que *chez*, *avec* ou le verbe *contenir*, se limite en général au contexte de la phrase. Or, il est clair que d'autres contextes, débordant le cadre de la phrase, sont susceptibles de donner accès à une relation d'hyperonymie. Par exemple, l'anaphore nominale infidèle en (103) permet d'établir que l'*Espace Visiteurs / Réunions / Formation* mentionné dans l'intertitre est un *équipement*.

(103) Espace Visiteurs / Réunions / Formation
Cet équipement donnera lieu à de nombreuses visites [Déplacements]

À la suite du travail mené par A. Condamines (2005) sur ces anaphores, nous avons voulu montrer la nécessité de prendre en compte le niveau du texte, d'une part en explorant la capacité d'une telle configuration à exprimer la relation recherchée, d'autre part, en

confrontant une non-spécialiste à des textes d'un domaine qui lui est inconnu [14]. Notre objectif était de mettre en évidence le fait que la restriction au contexte phrastique appauvrit l'analyse et qu'un parcours interprétatif guidé par les contextes permet de compléter une extraction de connaissances par patrons. Il s'agit de parcours interprétatif car la relation d'hyperonymie n'est pas aussi clairement exprimée par un tel contexte que par une phrase assertive telle que (104), reformulation libre de (103).

(104) L'espace Visiteurs / Réunions / Formation est un équipement qui donnera lieu à de nombreuses visites [exemple construit]

En (103), la rencontre du SN « cet équipement » est un **déclencheur** pour opérer une mise en relation plutôt qu'un marqueur de relation au sens où l'est la formulation en « est un ».

> la reprise anaphorique déclenche un phénomène plus ou moins conscient de recherche de référent (l'élément anaphorisé). Le locuteur peut ne pas aller jusqu'au bout de cette recherche mais le linguiste ou terminologue peut mettre en œuvre un processus de recherche d'un référent potentiel, qui lui paraîtra plus ou moins évident selon les contextes linguistiques et selon sa compétence dans le domaine. (Condamines & Jacques, 2006) [14]

Ce qui rend cette anaphore infidèle possible et ce qui lui permet de « fonctionner » effectivement dans un texte, c'est que la relation sur laquelle elle repose est présentée comme présupposée, et non assertée par le texte. Quand on a une succession telle que « Le chat a encore mangé une souris. Cet/L'animal est un chasseur redoutable. », la construction même invite à une mise en relation *chat / animal* qui sera d'autant plus facile si la relation est déjà connue, mais qui peut précisément être l'occasion de construire cette relation. Notre expérimentation s'appuie justement sur cette dimension instructionnelle de l'anaphore nominale infidèle.

Le dispositif expérimental mis en place a utilisé le corpus Vol Libre que j'ai présenté dans le chapitre 3. Je rappelle qu'il est constitué de divers articles parus dans la presse spécialisée du vol libre, dont le lectorat est constitué de pratiquants de l'activité. Il n'a donc pas été très étonnant de constater la rareté de contextes de définitions ou d'expression explicite de la relation d'hyperonymie. Le propos des textes n'est pas d'expliquer le domaine, mais de donner des appréciations sur le matériel commercialisé, le lecteur est supposé savoir de quoi l'on parle.

Notre propos était de tester, précisément dans ce genre de situation finalement assez courante dans les domaines spécialisés, dans laquelle les auteurs ne fournissent pas d'énoncés explicitant les relations, une méthode permettant tout de même de déduire les relations structurant le domaine. Nous avons ainsi cherché les contextes correspondant à une anaphore nominale infidèle : un SN déterminé par un démonstratif, dont le N n'est pas présent dans la même phrase ou dans la phrase précédente.

112 occurrences pour lesquelles on avait bien une relation d'hyperonymie ont livré 38 hyperonymes différents (listés dans le tableau 11).

Hyperonyme	Occ.	Hyperonyme	Occ.
aile	18	constructeur allemand	1
machine	16	débattement	1
catégorie d'aile	12	face	1
axe	10	importateur	1
configuration	5	marché germanique	1
catégorie	4	marque	1
constructeur	4	modèle Small	1
voile	4	nuages	1
manœuvre	3	parapente intermédiaire	1
style d'aile	3	phase	1
accessoire	2	phénomène	1
allure	2	position	1
modèle	2	principe	1
produit	2	qualités	1
vitesse	2	style de parapente	1
altitudes	1	style de voile	1
casquette	1	tissus	1
charge	1	valeur	1
choses	1	voile neuve	1

Tableau 11 : *Hyperonymes du corpus Vol Libre utilisés comme anaphore nominale infidèle*

Seuls 4 d'entre eux : *aile*, *machine*, *marque*, *produit*, sont aussi présents dans des contextes captés par les patrons lexico-syntaxiques « classiques ». La méthode semble ainsi un bon complément de la recherche d'énoncés riches en connaissance à base de patrons lexico-syntaxiques, mais est-elle vraiment exploitable pour qui ne connait pas le domaine ? Dans la démarche qui est celle de la terminologie textuelle (voir chapitre 3 et Bourigault & Slodzian, 1999), c'est en effet le texte qui est le premier informateur, il est donc essentiel de vérifier que, pour une recherche de relation qui ne s'appuie pas sur un contenu asserté mais sur une inférence, celle-ci soit réalisable à partir des éléments contenus dans le texte seul. C'est en fait ce que nous avons testé en premier lieu, en confiant à la néophyte (A. Condamines) le soin d'interpréter les contextes repérés et de proposer pour chacun le second terme de la relation.

Deux cas de figures principaux se sont présentés. Dans le premier, les contextes étaient facilement interprétables comme anaphore infidèle hyperonymique et, si l'on peut dire, tout allait bien sans plus d'efforts. L'extrait (105) en donne un exemple, l'hyperonyme est en gras, son antécédent est souligné.

(105) A bout d'accélérateur, ça déménage à presque 50 km/h, en butée de poulies. La voile reste solide, et j'avoue avoir abusé de **cet accessoire**, histoire de s'enlever plus vite du décor. [Vol Libre]

Dans le second cas de figure, il était nécessaire soit de chercher dans l'ensemble du corpus des indices pour l'interprétation, soit de consulter un expert. L'extrait (106), qui est un début de sous-section, donne un aperçu de ces contextes moins accessibles, le SN à interpréter est en gras.

(106) Les B, pour quoi faire ?
C'est vrai, j'en parle à tous les essais, alors que depuis quelques temps, cette pratique se marginalise, présentant vraiment peu d'intérêt en situation turbulente ou ventée. Et en situation calme, en avons-nous vraiment besoin ? Alors, dans rubrique essais à venir, je pense que **cette manœuvre** n'apparaitra plus. [Vol Libre]

De prime abord, l'analyste n'a pas su ce que désignait *cette manœuvre*. Une recherche de *manœuvre* dans l'ensemble du corpus donne accès à un descriptif plus explicite de ce que ce terme recouvre.

(107) Une **manœuvre** peu classique sur cette aile : les B. [Vol Libre]

(108) Toutes les **manœuvres** symétriques style B, frontales, sont gentilles, sans surprise ni excès dans les réactions. [Vol Libre]

(109) D'ailleurs, l'Arcus fut la préférée de toute l'équipe d'Aérogliss durant l'été, particulièrement lorsqu'il s'agissait de démontrer aux élèves certaines évolutions, style B, wing-overs radicaux, fermetures asymétriques et autres brusqueries. Elle reste homogène et compréhensive, sans aucun flou dans **ces manœuvres.** [Vol Libre]

(110) En route pour quelques vracs durant lesquels je laisse faire la voile. La frontale ? insignifiant ! La fermeture asymétrique à plus de 50 % ? pas même un quart de tour que tout est arrêté, ouvert et à plat. Trop gentil ! Les B ? Ah non, on en parle plus, c'est fini ! Notez seulement la douceur de l'abattée de sortie. Décidément, L'Epsilon se montre vraiment très sage... Et à défaut de thermiques, je répète **ces manœuvres**, confirmant ses qualités. [Vol Libre]

On voit resurgir dans ces divers contextes une partie de l'intertitre de l'extrait de départ, *les B*, avec d'autres termes, on peut donc en déduire que *manœuvre* est un hyperonyme pour « les B » – ainsi que pour (*fermeture*) *frontale*[107], *fermeture asymétrique*, *wing-overs*.

En fait, dans la plupart des cas, même si l'antécédent du SN démonstratif n'est pas immédiatement identifiable, même si la relation n'est pas facilement discernable, un parcours du corpus livre des indices qui permettent l'établissement du réseau de relations. L'accumulation des contextes produit la connaissance visée.

Il faut toutefois noter que, dans une situation authentique d'élaboration du réseau terminologique et notionnel d'un domaine, une telle démarche est sans doute trop couteuse en temps en ce qu'elle oblige à rechercher de nombreux contextes pour une décision. Mais cette étude nous a permis de mettre en évidence des structures discursives potentiellement informatives lorsque les marqueurs « classiques » de la relation d'hyperonymie font défaut et nous a permis de mettre l'accent sur la prise en compte du texte dans sa dynamique.

On a vu dans toute cette section une association étroite entre une description linguistique des phnénomènes et un outillage logiciel appuyé sur des techniques de TAL, description et outils mobilisés en tant qu'assistants pour la réalisation d'une tâche requérant *in fine* une

107 Dont l'utilisation en (110) montre un exemple de réduction de terme complexe, cf. section 3.1.2.2.

interprétation et une prise de décision d'un agent humain. Le traitement automatique ne supplante pas l'humain, il le supplée.

Le travail, encore une fois basé sur le repérage de marqueurs linguistiques et l'élaboration de patrons lexico-syntaxiques, que je présente maintenant s'inscrit dans la même philosophie.

5.3 Le corpus comme pourvoyeur de contextes pour aider la rédaction scientifique en anglais

Une application concrète de la description linguistique consiste à produire des ressources destinées à faciliter telle ou telle tâche, automatique ou humaine. J'ai évoqué dans la section 3.2 le travail mené autour du lexique scientifique transdisciplinaire : son inventaire puis le rangement des unités dans des catégories sémantiques (Hatier *et al.*, 2016) participent de la construction d'une ressource exploitable aussi bien pour le traitement automatique que pour l'enseignement de la langue[108].

Plus modestement, L. Hartwell et moi-même avons souhaité proposer une ressource appuyée sur Scientext[109], dans un objectif d'aide à la rédaction en anglais (Jacques, Hartwell & Falaise, 2013) [24]. Le cadre de l'étude est celui du « learning with corpora » (Boulton & Pérez-Paredes, 2014), c'est-à-dire l'utilisation de corpus pour développer des habiletés langagières, en concevant le corpus comme un *input* à partir duquel travailler certaines formes ou certaines constructions. J'y reviendrai plus particulièrement dans le chapitre 6, qui sera consacré aux applications didactiques.

Notre étude s'est fondée sur le constat de la nécessité pour tout scientifique de publier en langue anglaise, sinon un texte entier, au moins un résumé, et de la nécessité en conséquence pour le scientifique de maitriser un répertoire minimal de formes dans cette langue. Or peu d'entre nous avons la disponibilité (et l'offre de formation) pour travailler spécifiquement la rédaction en anglais. En outre, s'il existe des ouvrages ou des sites web destinés au public scientifique désireux d'améliorer sa communication écrite en anglais[110], ils ne peuvent offrir une vue extensive de contextes illustrant un même phénomène. De plus, bien souvent le besoin n'est pas de mieux comprendre les conditions d'emploi de telle ou telle unité lexicale en anglais, mais plutôt de savoir comment exprimer tel ou tel contenu. Cette équation, perspective onomasiologique et besoin de contextes, est typiquement celle à laquelle une base comme Scientext permet de répondre.

Scientext dispose d'un fonds conséquent d'articles scientifiques rédigés en anglais et offre, via l'interface Scienquest (Falaise, Tutin & Kraif, 2011), la possibilité de pré-coder des requêtes, des *grammaires locales* dans la terminologie de Scienquest. Notre étude a abouti à l'élaboration de nouvelles grammaires et, en collaboration avec le concepteur de Scienquest, A. Falaise, à un enrichissement des fonctionnalités de Scienquest dans les directions que je détaille maintenant.

La perspective onomasiologique implique de sélectionner un contenu à exprimer qui soit récurrent dans l'écrit scientifique et en particulier dans les résumés que les auteurs fournissent. Les travaux de Swales sur l'écrit scientifique (1990) et son modèle CARS (*Creating A*

108 Un ouvrage destiné à montrer les diverses exploitations du lexique scientifique transdisciplinaire est en préparation pour 2018 : Jacques & Tutin (en préparation). *D'une discipline à l'autre : lexique transversal et formules discursives des sciences humaines*. Éditions ISTE.

109 http://scientext.msh-alpes.fr/scientext-site/spip.php?article1

110 Voir par exemple http://users.wpi.edu/~nab/sci_eng/ page consultée le 21 aout 2017.

Research Space) mettent en évidence l'impératif, pour un auteur, de créer sa propre « niche » de recherche et donc de l'afficher clairement pour informer le lecteur et retenir son attention. La présentation des objectifs d'une recherche est ainsi un des éléments qui trouvent leur place aussi bien dans l'introduction d'un article scientifique, la section plus particulièrement étudiée par Swales, que dans le résumé, petit texte à fournir souvent en anglais même pour les articles écrits dans une langue autre que l'anglais, et sur lequel nous avions déjà commencé à travailler (Hartwell & Jacques, 2012) [23]. Cette présentation est cruciale car elle offre au lecteur la clé de la recherche exposée et lui permet de décider d'une lecture ou non de l'article. Pour ces deux raisons, nous avons pris comme objet la formulation de l'objectif de recherche, avec l'ambition d'en décrire les diverses expressions pour permettre à un rédacteur non-anglophone d'accéder grâce à Scientext / Scienquest à une variété de contextes. Nous retrouvions avec cette étude la démarche que j'ai décrite au chapitre 2 : partir d'une signification pour en décrire les réalisations discursives. Nous y avons ajouté l'élaboration de patrons lexico-syntaxiques, traduits en *grammaires locales* selon le « langage » de Scienquest.

Cette démarche s'articule en trois phases :

1. le recueil des données dans le corpus, c'est-à-dire des passages exprimant l'objectif de la recherche ;

2. la modélisation de ces passages, c'est-à-dire la transformation des formulations « de surface » en structures plus abstraites : les patrons lexico-syntaxiques ;

3. la formulation des grammaires correspondantes.

Les deux premières phases ont été menées de façon manuelle sur les résumés d'un sous-ensemble de 600 articles, provenant de revues de recherche variées afin de minimiser une éventuelle influence du support sur l'expression, et rédigés par des chercheurs d'universités anglaises, américaines et canadiennes, afin de capter autant que possible des auteurs anglophones natifs. La troisième phase a alterné, comme dans l'étude décrite en 5.2.1, élaboration de patrons et ajustements pour optimiser la précision[111]. Les patrons résultants sont ensuite évalués sur les résumés de l'ensemble de la base d'articles scientifiques en anglais.

Les formulations de l'objectif de recherche que nous avons relevées sont regroupées en 12 types, en fonction des patrons qui seront à construire pour les capter ensuite dans les textes. Ils sont illustrés par les extraits (111) à (122), dans lesquels les éléments qui servent à l'élaboration des patrons lexico-syntaxiques sont en caractères gras. J'insiste sur le fait que cette typologie ne se fonde pas sur une analyse linguistique en termes sémantique ou pragmatique mais est guidée par notre objectif applicatif, elle correspond donc aux contraintes en termes de lexique et de structure syntaxique applicables à chaque type. Néanmoins, ces contraintes sont des reflets des propriétés sémantiques et pragmatiques des formulations, on constate par exemple en position sujet des alternances régulières entre un pronom désignant le chercheur (*we*, *I*) et un SN désignant l'écrit ou la recherche (*the study, this article, this review,…*), qui constituent autant d'accroches formelles pour repérer automatiquement ces passages.

(111) **Here**, **we report** the results of a survey to assess the prevalence of drg in a globally representative panel of disease-associated meningococci.

(112) **Here**, a statistical **test** to detect gene conversion […] **is presented**.

111 Pour mémoire : la précision est le rapport entre le nombre de contextes pertinents et le nombre de contextes renvoyés par le système. Une précision de 88 % signifie que sur 100 contextes produits en réponse à la requête, 88 correspondent à ce que l'on cherchait.

(113) **In order to determine** which foods might be related to disease activity in UC a new method of dietary analysis was developed and applied.

(114) **To understand** the physiological processes responsible for elevated Cd accumulation in shoots and grain, Cd uptake and translocation were studied […]

(115) **In this** retrospective **review**, we examine whether progression to ESRF can be predicted and whether treatment [...]

(116) **In the current study** we report the isolation and preliminary characterization of homologous proteins from goat seminal plasma.

(117) **This article outlines** the evolution of a community pharmacy-based supervised consumption of methadone program in Grater Glasgow.

(118) **The present study addresses** the relationship of protein folding propensities to the evolutionary relationship between residues.

(119) **Our aim was to determine** the effects of taking a red clover-derived isoflavone supplement daily for 1 year on mammographic breast density.

(120) **We present** Homology Induction (HI), a new approach to inferring homology.

(121) **We aimed to assess** warfarin treatment in primary health care

(122) **This study** was **undertaken to characterize** the expression of chemokine receptors

Le dernier, (122), se modélise ainsi :

- un verbe d'étude ayant pour sujet un SN constitué avec un nom dénotant une recherche et pour complément un syntagme prépositionnel formé avec la préposition *to* ;

et est capté par la grammaire suivante :

```
$verb4=carry,conduct,undertake,perform,design   // liste verbes d'étude
$research=search,research,study,project         // liste noms de recherche
(SUJCOMP,#2,#1) = (SUJ,#3,#1) (AUX,#3,#2) ;  // sujet avec auxiliaire
Main = <form=$research,#2> && <lemma=$verb4,#1> && <lemma=to,#4> && <cat=/V/,#5> :: ((SUJ,#1,#2) OR (SUJCOMP,#2,#1)) AND (PREP,#1,#4) AND (NOMPREP,#4,#5) ;  // règle principale
```

Je reprends de (Jacques *et al.*, 2013 : 154) [24] le tableau 12 qui présente les mesures de précision sur le corpus initial de 600 articles et sur le corpus de test. Nous n'avons mesuré que la précision car le propos n'est pas de maximiser le rappel : il s'agit de donner à un rédacteur un aperçu contextualisé des formulations existantes, il est donc préférable de montrer les contextes les plus pertinents, quitte à avoir un peu de silence (c'est-à-dire des contextes pertinents non atteints).

Patrons	Corpus d'étude		Corpus de test	
	Nombre d'occ.	Précision	Nombre d'occ.	Précision
1 (Here, we...)	26	88,5 %	342	87,1 %
2 (Here, a test is...)	3	100 %	23	60,9 %
3 (In order to...)	12	66,7 %	104	86,5 %
4 (To V...)	60	36,7 %	393	77,1 %
5 (In this study)	34	88,2 %	421	88,6 %
6 (In the present study)	4	80 %	128	89,1 %
7 (This paper V)	94	81,9 %	498	80,5 %
8 (The present paper V)	12	75 %	156	87,8 %
9 (Our aim is to)	63	90,5 %	496	98,8 %
10 (We present)	154	77,9 %	500	87 %
11 (we V to)	213	63,8 %	259	82,6 %
12 (this study V to)	13	84,6 %	192	87,5 %

Tableau 12 : *Mesures de précision et nombre d'occurrences des douze patrons captant l'expression de l'objectif de recherche*

La précision globale moyenne calculée sur 3512 contextes (vérifiés par L. Hartwell qui est une locutrice native d'anglais) est de 86,5 %, cela signifie que 1 contexte sur 10 au moins n'est pas pertinent. Or notre public cible est constitué de locuteurs non natifs auxquels nous voulons présenter des exemples positifs. Cette ambition a conduit à intégrer dans Scienquest une nouvelle fonctionnalité : la possibilité d'éliminer certains contextes dans la fenêtre des résultats et la mémorisation des nouveaux résultats ainsi filtrés.

La figure 17 est une copie d'écran illustrant cette nouvelle fonctionnalité. La partie haute montre les contextes dont certains sont invalidés, la coche à leur gauche permet de les dé-sélectionner, la partie inférieure montre le choix donné à l'utilisateur soit d'exporter la sélection dans un autre logiciel, soit de la sauvegarder pour la retrouver dans une autre session, ce qui lui sera possible par la dernière commande « Restaurer des résultats préalablement sauvegardés ».

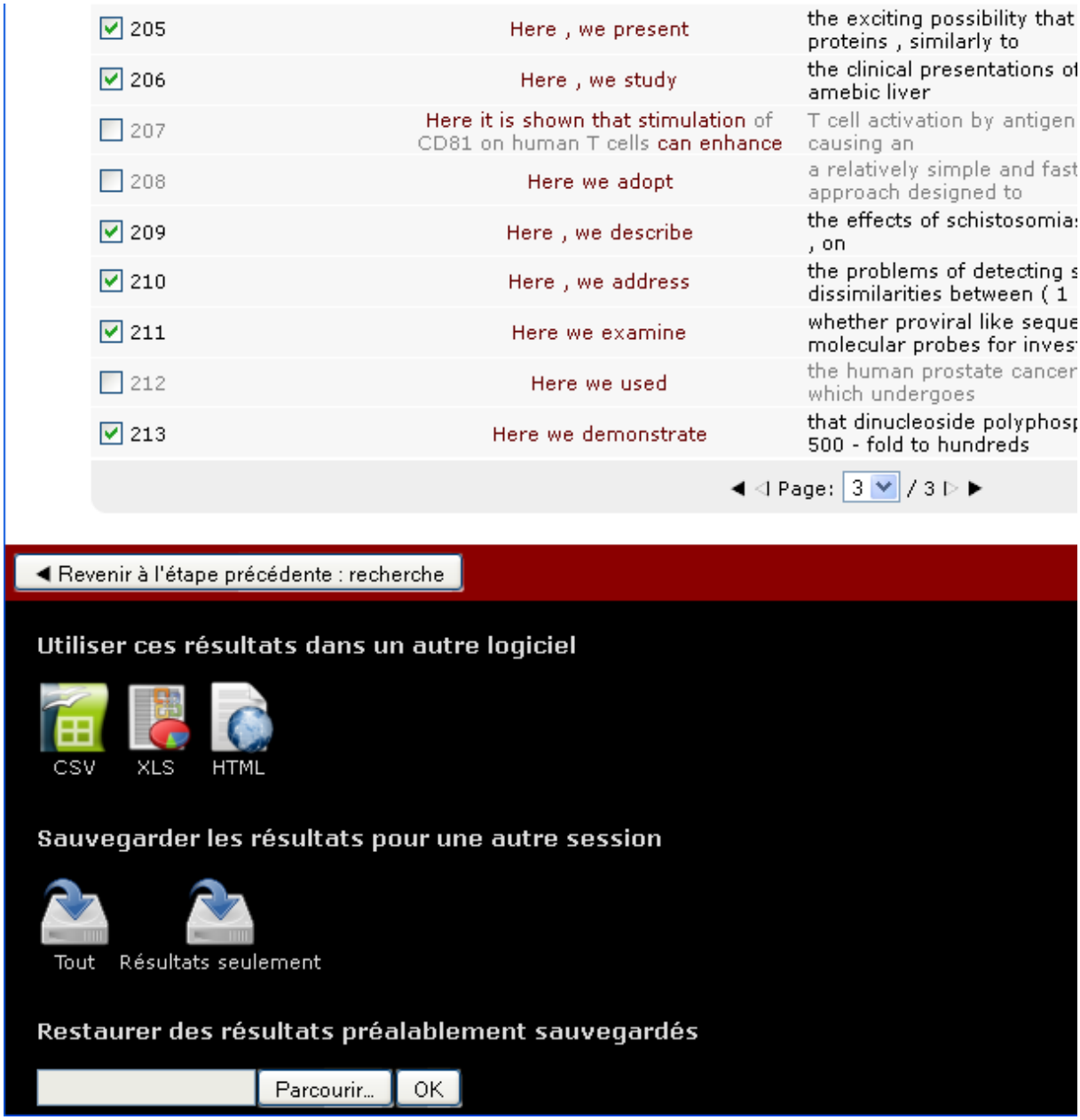


Figure 17 : *Fenêtre de sélection et sauvegarde des résultats d'une requête dans Scienquest*

Avec les requêtes pré-codées (les *grammaires* de Scienquest) et cette possibilité de sauvegarder une sélection des résultats, on fait d'un outil initialement dédié à la recherche linguistique un outil potentiel d'enseignement-apprentissage de la langue. Au passage, sont décrites certaines formulations sous un angle onomasiologique.

Cette étude mêle deux dimensions applicatives en exploitant les outils et procédures du traitement automatique des langues, pour la mise au point d'une ressource didactique. En ce sens, elle est emblématique des dernières orientations que prennent mes recherches et qui seront développées dans le dernier chapitre.

5.4 Synthèse-bilan du chapitre

Un premier axe applicatif a été exposé dans ce chapitre, ancré sur des tâches de traitement automatique. J'ai voulu y développer l'articulation entre la description linguistique et

l’application visée. Quelle que soit la tâche ultime, la démarche d’élaboration des traitements emprunte un même cheminement :

- circonscrire le phénomène linguistique à traiter ;

- en rassembler des occurrences à partir d’un corpus – par analyse manuelle ou par amorce automatique grossière ;

- en décrire les différents aspects en prenant en compte des caractéristiques pertinentes pour un traitement automatique – donc des traits formels et/ou calculables par un programme ;

- affiner les traitements par ajustements sur un corpus d’étude ;

- évaluer sur un corpus différent.

Les résultats des traitements permettent un retour sur le fonctionnement de la langue en mettant en évidence les aspects sur lesquels l’automatisation achoppe : dépendance au genre, nécessité d’interprétation...

L’application est ainsi un complément de la description car, même si elle s’appuie sur une description qui est orientée par le souci applicatif et qui de ce fait adopte une perspective spécifique, elle met en évidence certains fonctionnements linguistiques, certaines ambigüités ou zones de flou, ou elle accroit la connaissance des formulations possibles d’un contenu défini.

Chapitre 6 - Vers l'enseignement

Comprendre le fonctionnement de la langue pour son enseignement est précisément ce qui m'avait amenée à la linguistique (cf. chapitre 1). De ce fait, orienter les recherches linguistiques dans une telle direction est une évolution logique de mon parcours.

De la même manière que les traitements automatiques induisent une description attentive aux aspects formels, qui pourront être traités par l'ordinateur, la perspective de l'enseignement conduit à une attention aux phénomènes ou structures linguistiques qui constituent pour un apprenant autant de pierres d'achoppement. À un tel cadre doit répondre une double exigence :

- connaitre les particularités des compétences des apprenants et de leur développement, afin de cerner les acquis et les besoins ;

- définir précisément les objets à enseigner, ce qui implique d'être au clair sur leurs propriétés et les résistances qu'ils peuvent présenter.

Le chemin vers les applications didactiques ne sera achevé que si les chercheurs définissent ensuite des séquences ou des dispositifs d'apprentissage.

Ce dernier chapitre exposera des recherches encore en partie programmatiques qui visent à répondre aux enjeux de l'enseignement, en premier lieu de la langue maternelle, mais avec une petite incursion vers le Français Langue Étrangère. Les chapitres précédents ont montré mon ancrage sur la textualité, c'est donc assez naturellement la question des compétences textuelles qui constitue l'objet travaillé.

Dans l'objectif de documenter plus précisément ces compétences, un corpus a été constitué, que je présenterai en premier lieu (section 6.1). J'indiquerai au passage ses exploitations didactiques. Puis je tracerai dans une seconde partie de ce chapitre (section 6.2) les grandes lignes d'un projet afférent à ce champ.

6.1 Un corpus pour cerner les compétences rédactionnelles

J'ai participé à l'élaboration de plusieurs corpus à des fins de recherche, soit comme réalisation propre : [Vol Libre], soit lors de collaborations : [Déplacements], [Bricolage], [Médecine], [Termith], [EIIDA]. Toutefois, ces corpus sont restés pour l'essentiel confinés au sein de la petite équipe des chercheurs concernés, voire même dans mon seul ordinateur. Le corpus que je présente en 6.1.1 a au contraire été conçu dès l'origine[112] pour une mise à disposition de l'ensemble des chercheurs. Une première version en est disponible sur l'équipement d'excellence Ortolang « Outils et Ressources pour un Traitement Optimisé de la LANGue »[113].

6.1.1 « Littéracie Avancée » : des écrits académiques à un corpus

Produire des textes est à l'université une compétence incontournable. Les études supérieures impliquent pour l'étudiant d'accroitre son répertoire textuel pour y inclure des genres textuels plus ou moins nouveaux et surtout plus ou moins maitrisés. La plupart des cursus mettent l'étudiant en situation de produire des dossiers, des fiches de lecture, des rapports de stage, des projets en rapport avec sa discipline, des mémoires, etc., à côté des incontournables examens écrits qui sanctionnent les enseignements suivis, et pour ne parler que des écrits d'évaluation.

112 À l'initiative de Fanny Rinck, que je remercie de m'avoir associée à cette belle et stimulante aventure.

113 Site Ortolang : https://www.ortolang.fr/ (page consultée le 10 septembre 2017) ; page du corpus : https://www.ortolang.fr/market/corpora/litteracieavancee (page consultée le 10 septembre 2017).

Libersan *et al.* (2010) ont recueilli par enquête un inventaire de 138 genres académiques, dont certains sont en fait identiques ou presque sous des appellations différentes, par exemple en France au niveau Master, « mémoire de recherche » et « Travail d'Étude et de Recherches », ce qui réduit le nombre réel de genres distincts. Une telle profusion donne tout son sens à la notion de *littéracies universitaires* (Delcambre & Lahanier-Reuter, 2012), notion qui conjugue, à la fois dans son appellation et dans les recherches qui la prennent pour cadre, celle de *littéracie*, issue des travaux de J. Goody (par exemple 1979, 1994) et celle d'*écrit académique*. La multiplicité des genres et des exigences de l'écrit académique induit soit des transferts depuis d'autres genres travaillés tout au long de la scolarité, soit la construction de nouvelles habiletés langagières, et dans tous les cas implique de nouvelles compétences en matière rédactionnelle.

Comment mieux mesurer les acquis et les besoins de ces apprenants que sont les étudiants à tous les niveaux des cursus qu'en ayant un observatoire de leurs productions écrites ? C'est à cet objectif qu'entend répondre le corpus que nous avons nommé « Littéracie avancée » (Jacques & Rinck, 2017) [34]. Il rassemble des textes complets – même si ceux-ci ne constituent pas encore un document achevé – produits par des étudiants de divers niveaux de licence et master, dont le français est la langue maternelle – ou présumée telle, c'est-à-dire à l'exclusion des filières de FLE. Je vais d'abord synthétiser les ambitions de ce corpus et les principes qui ont présidé à sa constitution avant d'exposer, dans les sections 6.1.2 et 6.1.3, les premières exploitations auxquelles il a donné lieu.

6.1.1.1 Objectif majeur : être un observatoire des compétences du public étudiant

On le verra plus précisément à travers l'application décrite en 6.1.3, le corpus constitué s'inscrit dans une tradition qui associe recherche sur l'enseignement et corpus d'apprenants, dans l'esprit de (Cappeau & Roubaud, 2005) qui s'appuient sur l'analyse linguistique des textes d'élèves pour faire émerger et fonder les objets linguistiques à travailler pour le développement des diverses compétences à l'écrit. L'hypothèse sous-jacente est que l'arrivée à l'université ne doit pas signer la fin d'un enseignement / apprentissage de l'écrit mais doit au contraire favoriser le déploiement et la diversification de compétences rédactionnelles par un accompagnement continué[114].

Le principal objectif du corpus est de permettre des analyses des développements des diverses compétences et habiletés du public étudiant dans sa diversité, dans la même philosophie que les *learner corpora* (Granger, 2008) pour les apprenants de langues étrangères. Il doit permettre par exemple la comparaison de certains emplois lexicaux ou de certaines constructions (on en verra un exemple en 6.1.2), selon le genre produit ou selon le niveau universitaire.

Il se veut aussi un matériau potentiel pour la fabrication d'exercices ciblés, destinés à aiguiser le regard métalinguistique des apprenants en leur demandant par exemple de juger plusieurs versions d'un même contenu, ou plusieurs réalisations discursives d'une même intention communicative.

En un mot, nous l'avons conçu comme une **ressource**, aussi bien pour les chercheurs que pour les enseignants, ce qui implique divers choix méthodologiques pour sa constitution et sa mise à disposition. Pour notre cahier des charges, le corpus doit être évolutif, c'est-à-dire accepter d'intégrer de nouveaux ensembles de textes (car on le verra ci-dessous, il n'est pas constitué de

114 Les universités anglo-saxonnes et de nombreuses universités en Europe raisonnent sur ce modèle en institutionnalisant des *writing centers*, lieux dans lesquels les étudiants peuvent trouver un accompagnement personnalisé à l'écriture, voir http://writingcenters.org/

textes individuels mais de lots), il doit permettre la sélection de sous-ensembles pour la composition, par chaque chercheur, du sous-corpus pertinent pour sa recherche, sur le modèle de ce que permet Frantext[115] et après lui Scientext[116]. La conception et les métadonnées du corpus répondent à ces deux exigences.

6.1.1.2 Conception générale et métadonnées

Je l'ai dit, le corpus ne rassemble pas des textes envisagés individuellement mais des ensembles composés d'un minimum de dix textes. Cette contrainte répond à la conception sous-jacente de l'existence de « grammaires des genres » qui donnent forme à l'expression (même si, je l'ai évoqué au chapitre 5, il n'est pas si évident d'établir des traits et des fonctionnements linguistiques exclusifs d'un genre) et à l'ambition de saisir les productions selon ce critère de genre textuel. Dix productions inscrites dans un même genre et réalisées selon une même consigne peut sembler un chiffre trop réduit pour tirer des conclusions solides, mais c'est pour nous un compromis entre cette exigence d'avoir suffisamment de productions d'un même type et la considération pragmatique de pouvoir réaliser effectivement le corpus. Les sous-ensembles de dix textes au moins sont en effet rassemblés par des collègues universitaires, qui font signer aux auteurs des textes une autorisation d'exploitation, condition nécessaire pour la diffusion du corpus. Il n'est dans ce contexte pas toujours possible de réunir de gros volumes de productions.

Grâce aux collègues qui ont répondu à notre sollicitation, le corpus actuel comporte plus de 330 textes, environ un million de mots et une quinzaine de sous-ensembles distincts. Comme il rassemble des textes de genres et de niveaux divers, produits selon des consignes distinctes, il doit absolument s'accompagner des métadonnées qui à la fois renseignent sur ce que sont les textes et permettent des sélections de sous-ensembles, par exemple, tous les textes de M2, ou bien tous les mémoires, etc.

Chaque fichier est traité dans un format XML conforme à la TEI. Sont balisés toutes les structures textuelles (découpage en sections et paragraphes, intertitres, etc.) et certains éléments de mise en forme (gras, italique). Chaque fichier comporte un en-tête (*TeiHeader*) qui renseigne sur la fabrication du corpus (traitements, responsables, version, etc.) et inclut les métadonnées propres à chaque sous-corpus[117].

Chaque texte est ainsi accompagné d'une indication du niveau d'études de son auteur, codifié de façon simple en *L1 L2 L3*, *M1 M2*. De même, la « discipline » est explicitée, non de façon standardisée, mais en employant les expressions mêmes choisies par l'université pour intituler le diplôme. Par exemple, tel texte viendra du M1 « Formateurs d'enseignants 1er et 2nd degré », tel autre de L3 « Didactique du français ». Ces éléments sont complétés par le nom de l'université et l'année universitaire de production.

La langue des textes est aussi précisée, quoiqu'elle soit ici uniforme : français pour tous les textes. Un item est prévu pour la langue maternelle de l'auteur, mais jusqu'ici cette information n'a pas été consignée lors du recueil du texte et l'item reste donc non renseigné.

Plusieurs items concernent la nature du texte et ses conditions de réalisation. Le genre du texte est ici une donnée incontournable, mais qui pose évidemment de réels problèmes de définition. Il n'existe pas d'inventaire *a priori* et consensuel des genres académiques, bien que les

115 Frantext est une base textuelle développée à l'ATILF, largement utilisée par les linguistes français : http://www.frantext.fr/

116 http://scientext.msh-alpes.fr/ voir aussi le chapitre 4.

117 Je reprends de (Jacques & Rinck, 2017 : 225) [34] le descriptif de ces métadonnées.

curricula comportent des productions écrites peu ou prou similaires : quelle que soit la discipline, un étudiant peut s'attendre à réaliser une *fiche de lecture*, un *rapport de stage*, un *dossier*, un *mémoire*... Même dans le cas d'une dénomination commune, les exigences quant à la longueur, au contenu et à la teneur de l'écrit attendu sont variées. La solution la plus transparente consiste là encore à inscrire comme genre la dénomination produite par l'institution elle-même. On trouve donc dans le corpus des *compte-rendus de lecture*, des *synthèses théoriques* à côté de *mémoires* et *compte-rendus professionnels.* Il s'agit bien du genre prescrit et non d'une caractérisation *a posteriori* du texte en fonction de propriétés qui le feraient appartenir à tel ou tel genre (Malrieu & Rastier, 2001). Des études peuvent alors être menées sur les propriétés formelles d'un genre tel que les étudiants se l'approprient et sur leurs difficultés par rapport aux propriétés attendues.

En complément de cette indication de genre, est reportée la consigne d'écriture du texte. Il s'agit là encore d'une information que les chercheurs peuvent mobiliser soit comme préalable à une étude par exemple contrastive d'un même genre mais de consignes variées, soit comme variable explicative des observations qui peuvent être menées sur le-s texte-s.

De même, avec chaque texte est précisé le nombre d'auteurs. Peu de textes sont écrits « à plusieurs mains », dans la mesure où les cursus universitaires s'appuient massivement sur les productions individuelles pour la diplomation, huit seulement sont le fait de deux auteurs, mais ils peuvent amorcer des études contrastives pour mesurer les effets de ce facteur.

Deux derniers items liés aux conditions de production concernent le nombre de versions et le numéro de version du texte. Certains textes sont en effet une réécriture d'une version précédente. Les utilisateurs du corpus peuvent ainsi s'ils le souhaitent interroger ces textes via un alignement de ces versions.

Ajoutons à ce qui précède que la taille en nombre de mots est indiquée pour chaque texte, et l'on a un tableau complet des éléments qui sont à la fois supposés pertinents et informatifs et réalistement récoltables.

Ces éléments ont été pensés pour permettre des sélections par un outil d'exploration de corpus tel que TXM[118] (Heiden *et al.*, 2010), qui autorise de bâtir autant de sous-corpus que souhaité à partir d'un même corpus de départ. Ils ont été déterminés en fonction des utilisations que nous envisagions pour ce corpus, qui vont au-delà de nos propres sphères de recherche.

6.1.1.3 Études programmées par le corpus

Je l'ai indiqué précédemment, faire un corpus pour le mettre à disposition d'un large public de chercheurs suppose une démarche différente de la fabrication d'un corpus pour ses propres recherches, une démarche qui se décentre de ses thèmes de recherche et envisage le corpus par rapport aux besoins et aux intérêts d'un champ de recherches.

Clairement, le cadre dans lequel notre démarche s'inscrit croise la question de l'écriture au niveau universitaire et celle de l'enseignement de la langue maternelle. C'est donc sous le double angle de la production textuelle achevée avec ce qu'elle montre des compétences des scripteurs et de la ressource pour l'enseignement que le corpus a été pensé.

Le corpus Littéracie Avancée peut servir de support à une diversité d'études empiriques, qu'elles portent sur l'usage d'un mot, comme par exemple la conjonction *car* et son insertion dans le texte (Masseron, 2004), sur un phénomène textuel comme l'anaphore (Boch & Rinck,

118 http://textometrie.ens-lyon.fr/?lang=fr [page consultée le 15 septembre 2017]

2016), ou sur le positionnement énonciatif comme la citation (Kara, 2004)... Sur un plan plus formel, les acquis orthographiques et les zones qui manifestent la résistance de la langue écrite pourront être scrutées (Jacques, 2016a) [29] – par exemple, une étude exploratoire non encore publiée montre la quasi-absence d'erreurs d'orthographe lexicale mais le maintien de nombreuses erreurs sur la morphographie. La présence, dans le corpus, d'écrits élaborés aux différents niveaux d'études, permet de dégager des tendances pour le développement des diverses compétences en jeu. La focalisation sur un niveau et un genre précis pourra être privilégiée à des fins de comparaison avec d'autres corpus, particulièrement pour le FLE (Français Langue Étrangère) ou le FOU (Français sur Objectifs Universitaires), qui manquent souvent de données sur ce que les natifs eux-mêmes sont en mesure de produire aux différents niveaux d'un cursus universitaire ; j'en donnerai un exemple en 6.1.2.

Le corpus est aussi destiné à répondre à des besoins concernant l'enseignement. Deux perspectives non exclusives nous semblent possibles : dégager des études empiriques un inventaire des besoins des étudiants en matière de compétences à acquérir ; utiliser le corpus comme matériau pour la fabrication de ressources didactiques par exemple sous forme d'exercices et/ou d'objets d'observation appropriés.

> Les textes du corpus mettent en évidence réussites et maladresses de leurs auteurs. Les réussites ont ceci de remarquable qu'elles permettent de restituer aux apprenants des formulations, des passages qui ne présentent pas le caractère souvent ressenti comme inaccessible de l'écriture experte, mais des solutions à leur portée, élaborées par leurs pairs, qui peuvent alors jouer pleinement le rôle de *modèle*. Le contraste avec des passages similaires non réussis (par exemple des extraits d'introduction, des passages censément explicatifs ou argumentatifs), le recours à des jugements d'acceptabilité alimentent une réflexion sur ce qui fait que « ça fonctionne » ou « ne fonctionne pas ». (Jacques & Rinck, 2017 : 221) [34]

On verra avec l'exemple donné dans la section 6.1.3 comment une partie du corpus a été exploitée pour le développement d'une posture réflexive à travers des exercices en-ligne.

On pourra trouver dans (Jacques & Rinck, 2017) [34] d'autres pistes d'exploitation, les deux sections qui suivent vont donner du corps à ces propositions.

6.1.2 Ébauche de comparaison : étudiants français / étudiants chinois

Le travail que j'évoque maintenant n'a pas encore donné lieu à publication – celle-ci est en préparation, suite à une communication au 9e Colloque International de Linguistique de Corpus[119]. L'étude, menée avec Rui Yan, incarne le versant « aperçu des compétences rédactionnelles » que j'ai évoqué dans la section précédente et met en évidence l'intérêt d'un corpus tel que *Littéracie avancée* comme informateur sur les compétences des natifs. Elle vise à comparer dans deux corpus de mémoires d'étudiants, de langue maternelle française et apprenants chinois du français, l'utilisation de la structure [comme ... Verbe], dont (123) et (124) donnent des exemples (je mets en caractères gras la structure concernée).

(123) **Comme Gérard Chauveau le souligne**, le mot *lire* admet une polysémie redoutable [Littéracie avancée]

(124) **Comme je l'ai sous-entendu** précédemment l'enseignante joue également un rôle important dans ces échanges. [Littéracie avancée]

119 https://cilc2017.sciencesconf.org/?forward-action=index&forward-controller=index&lang=fr [page consultée le 15 septembre 2017]

Avant de donner des éléments sur ces emplois dans les deux populations, les grands traits des fonctions de cette construction, telle qu'elle a été décrite dans l'écrit scientifique, éclairent l'intérêt que nous lui portons.

Ainsi que je l'ai évoqué dans le chapitre 3, section 3.2, l'écrit scientifique mobilise entre autres un lexique spécifique, qui dans les textes s'insère dans une phraséologie spécifique. La structure [comme ... Verbe] est ainsi très productive dans l'écrit scientifique (Grossmann, 2014) où elle remplit un double rôle. Associée par exemple à un verbe de constat, tel que *constater*, *voir*, *observer*, *remarquer*, elle participe de l'inclusion dans le propos du lecteur pour lequel elle opère un soulignement de ce qui est affirmé conjointement. Son caractère quasi détaché, qui lui permet d'adjoindre une information supplémentaire à la proposition principale, contribue à « [dissocier] clairement le fait introduit et la prise à témoin du lecteur » (Grossmann & Tutin, 2010 : 6). C'est cette fonction qui est présente dans (123), avec non pas un verbe de constat, mais un recours à une autorité extérieure, sous l'ombrelle de laquelle s'abrite le propos qui suit. Elle joue ainsi « un rôle rhétorique de renforcement de l'argumentation » (Grossmann, 2014 : 764). Ce n'est toutefois pas là sa seule fonction. Quand elle s'appuie en particulier sur le verbe *voir*, la structure assure la cohésion textuelle en liant l'énoncé qu'elle accompagne à d'autres « lieux » du texte ; « l'auteur signale au lecteur qu'il a déjà traité du thème ou qu'il a l'intention de l'examiner plus loin » (Chambers, 2010 : 14). Cette seconde fonction est présente, maladroitement, dans (124). Elle offre la possibilité à l'auteur de revenir à un propos antérieur sans paraitre se répéter (125), ou d'anticiper sur un développement à venir (126).

(125) **Comme nous l'avons déjà dit dans l'introduction**, c'est seulement le phénomène d'activation d'une facette conceptuelle de l'entité nominale et non l'activation d'un aspect informationnel de l'entité relationnelle (verbe ou adjectif) qui sera traité ici[120]. [Termith – linguistique]

(126) **Comme nous l'argumenterons dans la suite de cet article,** nous sommes en désaccord avec cette affirmation de G. Kleiber[121] [Termith – linguistique]

En résumé, cette structure remplit une fonction argumentative et une fonction métatextuelle assez typiques de l'écrit scientifique.

La maitrise de telles constructions est donc un enjeu pour les étudiants engagés dans l'écriture scientifique, aussi bien natifs que non-natifs. Comme je l'ai souligné précédemment, pour les natifs, les genres textuels propres à cette écriture, tels que mémoire de recherche ou TER[122], sont à maitriser, pour les non-natifs, à la maitrise de ces genres textuels s'ajoute celle de la langue elle-même. Les questions qui sous-tendent notre étude sont très similaires à celles qui sont traitées par (Granger & Paquot, 2009) : avec quels verbes cette structure apparait-elle, dans les écrits des étudiants, natifs comme non-natifs, et dans ceux des experts – les experts étant les auteurs scientifiques publiés ? Y a-t-il des différences d'emploi entre étudiants et experts et entre étudiants natifs et non-natifs ? La comparaison doit permettre de comprendre la part de difficultés inhérentes au maniement de nouvelles formes discursives et celles inhérentes à la langue étrangère ou seconde.

120 Gamallo Otero, Pablo (2000). L'interprétation d'expressions nominales analysée comme un processus métonymique. *Revue de Sémantique et Pragmatique*, 7, pp. 29–58.

121 Godart-Wendling, Béatrice (2000). Comment ça réfère ? *Revue de Sémantique et Pragmatique*, 7, pp. 105-122.

122 Travail d'Étude et de Recherches.

Trois corpus ont été explorés, à l'aide de l'outil TXM :

- les articles de linguistique de Termith (cf. chapitre 3), représentant environ 450000 mots ;

- les mémoires de Master de Littéracie Avancée, représentant environ 460000 mots ;

- des mémoires de master rédigés en français par des apprenants chinois, dans diverses disciplines (linguistique, littérature, traduction, socio-culturel), représentant environ 600000 mots.

On le voit, deux biais éventuels incitent à la prudence quant aux résultats. Le dernier corpus couvre plusieurs disciplines, mais comme il reste tout de même dans le champ des sciences humaines, nous considérons qu'il permettra néanmoins de dégager des tendances. En revanche, le premier corpus est constitué d'articles, donc de textes plus courts, ce qui crée le risque d'observer chez les experts surtout des [comme ... Verbe] à fonction argumentative et moins à fonction métatextuelle – Grosmann et Tutin (2010 : 8) notent que les emplois cohésifs de *voir* sont presque deux fois plus nombreux dans les thèses que dans les articles. Nous garderons cette réserve à l'esprit pour l'analyse.

Pour répondre à la première question, nous avons systématiquement cherché dans les trois corpus les occurrences de la structure et avons éliminé celles qui sont en fait des comparaisons, par exemple (127) :

(127) La mélodie des structures poétiques est appréciée, reconnue et provoque un attrait pour les élèves **comme les comptines qu'ils apprennent plus tôt**. [Littéracie avancée]

Ont ainsi été recueillies 165 occurrences pour les experts, environ 300 pour les natifs et 174 pour les non-natifs. Dans la mesure où les longueurs des écrits des étudiant sont comparables, il apparait déjà que les étudiants natifs recourent bien plus volontiers à cette structure que leur homologues chinois. Le tableau 13 indique les 10 verbes les plus fréquents dans la structure pour les trois catégories de scripteurs.

Experts		**Natifs**		**Non-natifs**	
montrer	32	voir	72	dire	64
voir	28	souligner	45	savoir	19
dire	9	expliquer	32	écrire	14
indiquer	8	dire	31	voir	9
souligner	8	évoquer	13	illustrer	5
témoigner	8	préconiser	12	remarquer	5
savoir	7	préciser	11	analyser	4
signaler	7	remarquer	9	indiquer	4
démontrer	6	montrer	8	mentionner	4
proposer	6	signaler	6	parler	4

Tableau 13 : *Dix verbes les plus fréquents dans chaque corpus*

Il est remarquable d'observer une telle disparité, à la fois en ce qui concerne les types et en ce qui concerne les fréquences d'emploi. Un seul verbe est présent dans les trois corpus, le verbe

voir, mais il n'occupe pas la première place chez les experts alors que pour les étudiants français, il écrase littéralement le reste du tableau. Chez les non-natifs, le verbe *dire* qui occupe la première place est bien représenté aussi chez les natifs où il constitue 10 % des occurrences, mais reste très marginal chez les experts puisqu'il n'y représente que 5 % des occurrences. On notera aussi la présence du verbe *parler* chez les non-natifs, sur lequel je reviendrai plus loin.

Un dernier élément chiffré concerne le nombre de verbes différents : il est en fait assez comparable puisque les experts recourent à 36 verbes différents, les natifs à 41 verbes différents et les non-natifs aussi à 36.

Comment cette structure est-elle utilisée, semble-t-elle maitrisée par les étudiants ?

Pour l'examen de la fonction privilégiée, je comparerai uniquement les écrits des étudiants, puisque, comme je l'ai signalé, les différences de taille des textes concernés introduit un biais par rapport aux experts. Le tableau 14 montre pour chaque fonction le nombre d'occurrences et le pourcentage qu'elles représentent.

	Natifs		**Non-natifs**	
Fonction	occurrences	pourcentage	occurrences	pourcentage
argumentative	197	65 %	132	75 %
métatextuelle	103	35 %	42	25 %
Total	300	100 %	174	100 %

Tableau 14 : *Répartition des usages de la structure selon sa fonction*

Ces chiffres montrent que la fonction métatextuelle semble moins assimilée par les non-natifs que par les natifs. Mais il ne faut pas s'en tenir aux chiffres, la caractéristique la plus frappante des usages métatextuels de la structure est qu'elle fonctionne quasi-exclusivement à rebours, en rappelant des thèmes déjà abordés. Seules 6 occurrences, 4 dans le corpus de natifs, 2 pour les non-natifs, annoncent des développements ultérieurs :

(128) **Comme nous allons le voir**, la réécriture permet de développer un style personnel [Littéracie avancée]

(129) En effet, **comme nous le verrons plus tard**, l'activité de réemploi lexicale qui a été faite après et sur le thème des animaux de la savane. [Littéracie avancée]

(130) **Comme on va le voir ci-dessous**, la multiplexité fait fragile la consécution [Sup-chinois]

Les extraits (129) et (130) en témoignent, la conscience du fait que la gestion du texte peut s'opérer aussi par anticipation n'exclut pas certaines maladresses d'expression. Et c'est particulièrement ce qui manifeste la fragilité des acquis en matière rédactionnelle, y compris à un niveau d'études supérieures.

L'examen complémentaire des occurrences manifestant une fonction plus argumentative met en évidence d'autres caractéristiques. Dans les deux corpus, on observe un recours assez massif à une autorité extérieure avec des tournures qui introduisent un discours rapporté – alors que les experts usent plus volontiers de reformulation (Boch, 2013 : 560) – ce qui est congruent avec les études antérieures sur les écrits d'étudiants. Que ce soit pour la fonction métatextuelle ou pour l'argumentation, les non-natifs sur-emploient des verbes de parole tels que « dire », qui arrive en première position dans le tableau 13, ou « parler », ex. (131) et (132).

(131) **Comme nous l'avons déjà parlé**, le goût pour l'artificialité avait pour origine l'opposition d'une société industrielle [Sup-chinois]

(132) L'histoire dans La Modification est très simple, **comme nous avons parlé**, puisqu'il s'agit seulement d'un homme qui fait un voyage en train de Paris à Rome. [Sup-chinois]

Ces quelques extraits le montrent, pour certains étudiants, aussi bien natifs que non-natifs, se révèlent deux difficultés majeures. Puisque la construction est largement employée pour introduire des propos rapportés, elle suppose un choix de verbe de dire. Mais la norme en français écrit est de varier le lexique autant que faire se peut pour éviter les répétitions. Les scripteurs sont ici en tension entre la nécessité de marquer le discours rapporté et celle de ne pas s'en tenir à une seule forme verbale pour ce faire. On voit ainsi surgir des emplois de verbes totalement inappropriés, tels que *avouer* ou *dénoncer*, ex. (133), (134).

(133) le père et les enfants [jouent] des rôles dépendants **comme l'auteur l'a avoué** dans les Mémoires intérieurs que « notre enfance nous apparait comme une nébuleuse dont une mère est le noyau tendre et rayonnant ». [Sup-chinois]

(134) Mais **comme le dénonce Manesse** (2007 : 211) « Peut-on en effet exiger des adolescents d'aujourd'hui qu'ils maitrisent aussi bien la langue qu'autrefois alors qu'ils disposent de moins de temps pour la mémoriser et qu'ils ont davantage de matières à étudier ? ». [Littéracie avancée]

Ces exemples mettent en évidence une absence de prise en compte ou même de perception de la connotation qui devrait réserver l'usage de ces verbes à des contextes autres.

Le deuxième problème, lui aussi lié à une question de registre, tient à l'irruption de formulations plutôt familières et orales dans ces écrits.

(135) L'enseignant, **comme dit plus haut**, et le garant du langage (sic) [Littéracie avancée]

(136) **Comme je le présente dans des paragraphes d'avant**, Hugo est bien économique, même un peu avare en quelque sorte. [Sup-chinois]

Pour les non-natifs, s'ajoutent des difficultés liées à la connaissance de la langue. Il n'utilisent jamais la tournure avec des sujets inanimés – des objets du texte lui-même : tableaux, figures, exemples – comme s'ils ne la percevaient que dans sa capacité à introduire une citation directe. Dans cette dernière fonction, la syntaxe autour du verbe est parfois malmenée : peu d'inversion du sujet là où elle « ferait plus naturel », placement des clitiques erroné, par ex.

(137) Ce thème, **comme ce que le poète a dit lui-même** : « … est une sorte d'autobiographie poétique … » [Sup-chinois]

Ces diverses erreurs, aussi bien des natifs que des non-natifs, dessinent des champs d'intervention possibles, telles que celle qui sera présentée en 6.1.3.

Cette étude, encore dans sa phase initiale, montre le bénéfice qui peut être tiré de l'analyse contrastive de corpus d'écrits d'étudiants. Elle a pour nous mis en évidence que certaines difficultés rencontrées par les non-natifs ne leur sont finalement pas propres et sont au bout du compte moins liées à la langue qu'à la maitrise d'une phraséologie typique des genres. La gestion du texte, de l'introduction des propos rapportés sont autant de problèmes d'écriture à résoudre, pour les non-natifs comme pour les natifs.

C’est pour ces derniers qu’a été conçue l’application didactique que je vais maintenant brièvement évoquer, en signalant qu’elle n’a pas non plus fait l’objet d’une publication, mais de plusieurs présentations orales[123].

6.1.3 Une application pour aider la rédaction

L’application que je présente ici a reçu un financement dans le cadre d’un appel à projet PedagoTice de l’Université Joseph Fourier (Grenoble) en 2014-2015 et a été réalisée en collaboration avec Fanny Rinck, qui m’a accompagnée dans mes élans informatiques. Pour faire saisir la philosophie qui nous a guidées, je reprendrai ce que nous avons écrit ensemble :

> Il ne s'agit pas de demander à l'apprenant de simplement « trouver la réponse juste », mais d'aiguiser son regard et de l’entrainer ainsi à une réflexivité essentielle dans le développement de ses compétences scripturales. (Jacques & Rinck, 2017 : 221) [34]

Les présupposés de la démarche sont multiples. D’abord, on considère que les compétences de rédaction à l’écrit peuvent se travailler à travers des dispositifs dédiés, à tous niveaux de la scolarité. En clair, l’arrivée dans l’enseignement supérieur ne devrait pas marquer la fin d’un travail formel sur l’acquisition des compétences d’écriture mais devrait en proposer une continuation. Ensuite, l’angle d’attaque choisi est celui des habiletés linguistiques. On fait l’hypothèse qu’en travaillant la matière du texte, en s’interrogeant sur des solutions linguistiques à trouver pour les problèmes d’écriture, c’est l’intelligibilité du texte qui se travaille et c’est le raisonnement qui se construit. Enfin, ainsi que l’indique la citation qui précède, on défend l’idée d’une réflexivité (Boutet, 2005) à base d’une observation et de manipulations proches de celles de l’analyse linguistique, en considérant que la conscience linguistique constitue un outil stratégique pour le contrôle et la révision lors de la production de textes.

Le propos du programme PedagoTice, qui a financé le projet, est de favoriser la mise en place d’enseignements appuyés sur les TICE (Technologies de l’Information et de la Communication pour l’Enseignement). C’est donc à travers un cours en ligne, hébergé par une plate-forme pédagogique, que nous avons élaboré des exercices.

Comment répondre aux besoins des étudiants, les mettre en situation d’aiguiser leur regard sur des problèmes authentiques et par là de prendre conscience du fonctionnement de la langue et des textes ? Trois temps ont permis la réalisation du projet.

1. Relevés des erreurs dans les textes

Une partie seulement du corpus Littéracie avancée a été exploitée, celle qui est constituée d’écrits courts (3 à 4 pages) préparatoires au Concours de Recrutement de Professeur des Écoles (CRPE). Les étudiants de master préparant ce concours constituent notre cible, nous voulions les faire travailler dans le genre à produire – nos travaux respectifs, aussi bien pour Fanny Rinck que pour moi-même, mettant en évidence l’importance du genre pour la construction du texte. Et comme ce sont des écrits que nous corrigeons depuis plusieurs années pour la validation universitaire, nous avions déjà une perception des besoins et des types d’erreurs. La première phase a consisté à relever les erreurs dans chacun des types préétablis,

123 Jacques M.-P. & Rinck F. (2015). Une linguistique fondamentale et appliquée à base de corpus. *TRELA (Terrains de Recherche en Linguistique Appliquée)*, Paris.

Jacques M.-P. & Rinck F. (2015). Compétences rédactionnelles à l'université : un corpus et un projet PedagoTice. *Journée d’étude Corpus et Didactique*, Valence.

qui allaient des problèmes d'orthographe grammaticale à la structuration d'une introduction (très codifiée dans ce genre) en passant par les choix lexicaux, la syntaxe, en particulier verbale, la gestion des anaphores, la progression au sein d'un paragraphe, la gestion de l'argumentation, etc.

2. Choix des types d'exercices

Des dizaines d'extraits « en fichier », qu'en faire, comment les transformer en exercices susceptibles de répondre à nos objectifs ? Il ne s'agit pas là d'une interrogation technique mais bien didactique. Avec la contrainte de l'impossibilité de demander une réécriture, qui ne peut être évaluée automatiquement, la question se pose du meilleur type d'exercice à proposer en distanciel pour chaque type d'erreur. Par exemple, pour travailler l'orthographe, on peut aussi bien demander de choisir dans une liste la graphie correcte, que placer un « trou » au lieu des lettres finales et demander d'écrire la terminaison du mot dans le trou, ou encore demander de repérer dans une phrase les mots mal orthographiés (il existe sans doute encore d'autres possibilités, je m'en tiens à celles que permettent les plates-formes disponibles dans notre université).

Les questions de type « texte à trous » mobilisent des connaissances effectives : il s'agit de choisir ou de définir la graphie correcte, le mot approprié, la bonne expression, la bonne ponctuation… Les questions de types QCM mobilisent plutôt des compétences réflexives : il s'agit d'identifier, dans un ensemble donné, les formulations, graphies, choix lexicaux, etc. qui répondent à une certaine norme écrite. On fait l'hypothèse que ces deux types de connaissances et compétences forment la compétence rédactionnelle. Les exercices ont donc été élaborés en alternant ces deux perspectives. Les figures 18 et 19 donnent des exemples d'exercices de diverses sortes.

La plate-forme utilisée, Chamilo[124], permet de regrouper les exercices en *parcours*. La figure 20 montre les différents parcours proposés dans le cours.

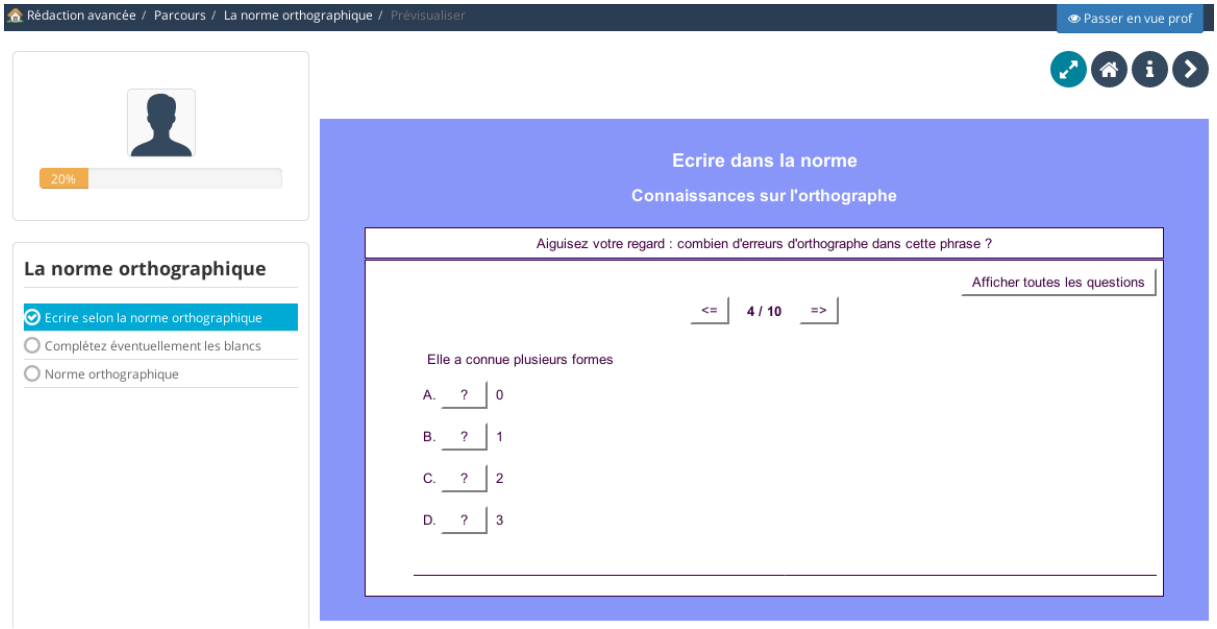


Figure 18 : *un exercice orthographique réflexif*

124 https://chamilo.org/fr/

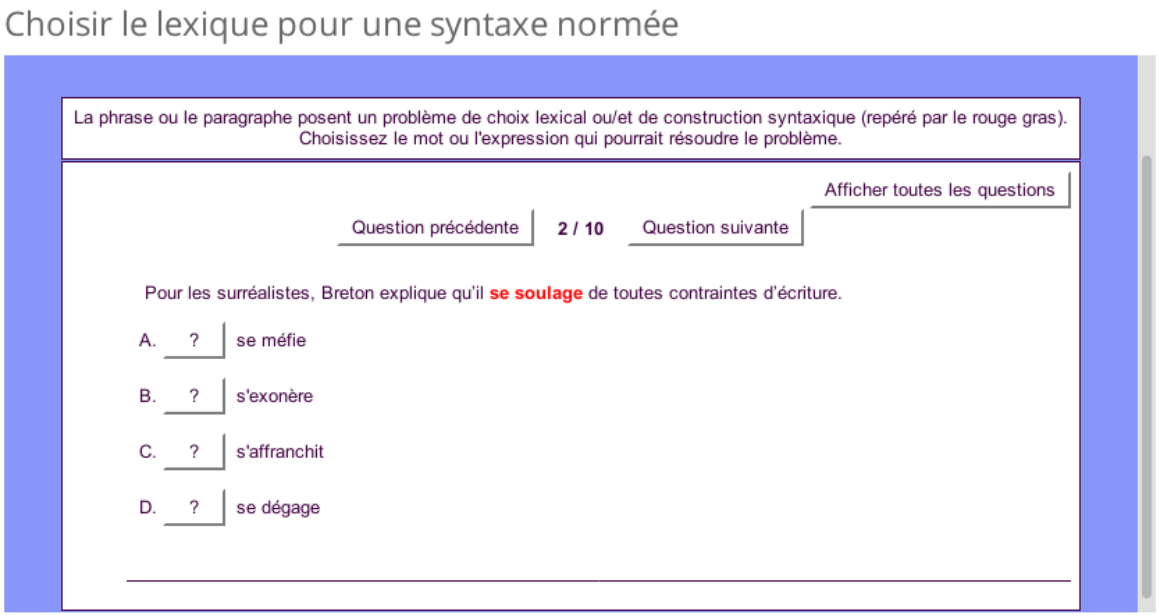


Figure 19 : *un exercice de connaissance lexicale (plusieurs réponses sont possibles)*

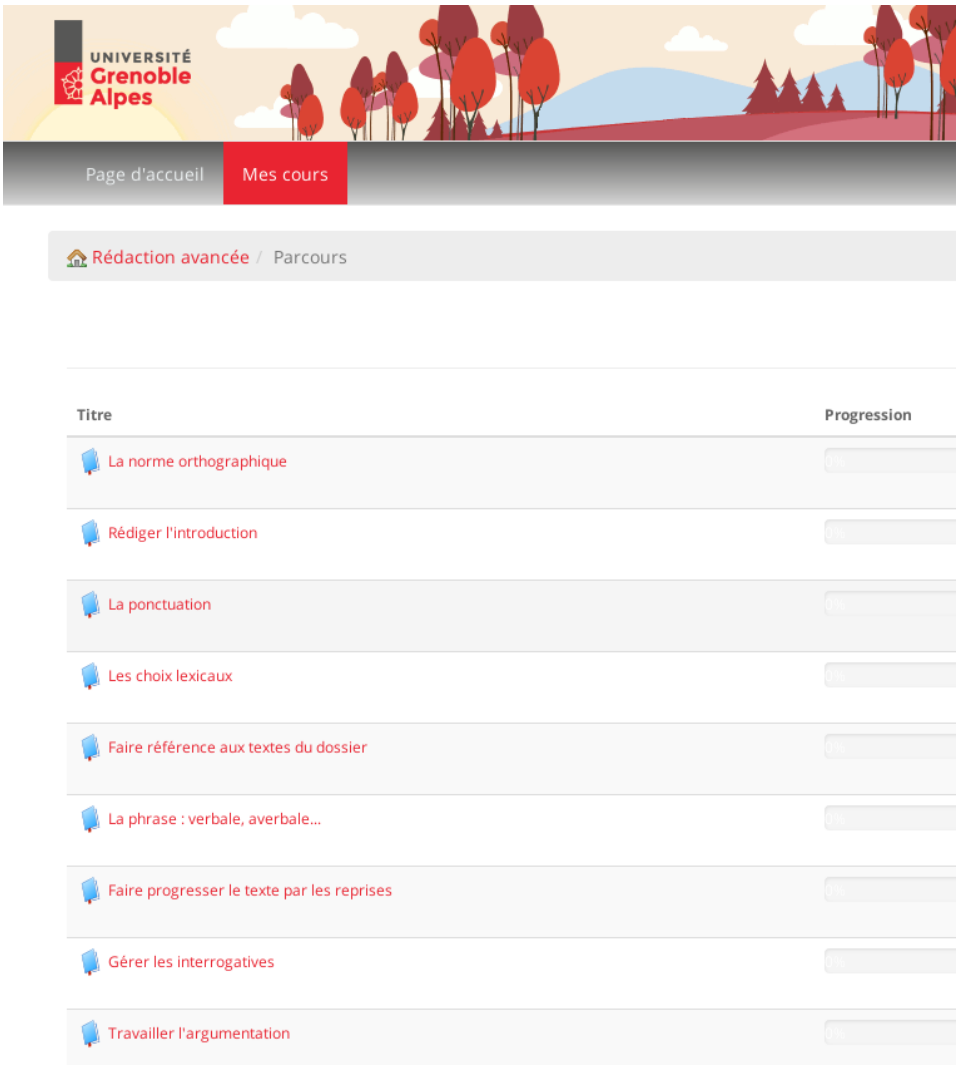


Figure 20 : *Les parcours du cours « Rédaction avancée » bâti sur Littéracie avancée*

Après l'utilisation de ce cours, la troisième étape est celle du bilan.

3. Bilan : utilisation du cours par les étudiants

La plate-forme permet d'accéder à des statistiques d'utilisation, mais, malheureusement, nous n'avons pu collecter des statistiques de réussite ou de progression. Indépendamment des progrès ou non des étudiants, que de toute façon il serait illusoire de penser mesurer à l'échelle d'une année universitaire, nous nous sommes intéressées aux « parcours » (c'est-à-dire regroupements d'exercices) qui ont été les plus fréquentés et, *a contrario*, les moins fréquentés, en tant que révélateurs des besoins ressentis par les étudiants.

Ont été massivement travaillés les parcours suivants :

1. Travail de l'argumentation
2. Choix lexicaux
3. Orthographe
4. Ponctuation

En revanche, ont été « boudés » les parcours :

1. Travail sur les phrases
2. Progression du texte par les reprises

Les étudiants semblent donc, par leur choix, penser qu'ils n'ont pas à faire de travail particulier sur la construction normée des phrases, pas plus que sur la progression du texte à travers la gestion des anaphores. Comme leur production dément la maitrise de ces aspects des textes, on peut en tirer la conclusion qu'il faut d'abord agir sur la prise de conscience de ces besoins.

La réalisation de ce cours laisse néanmoins entrevoir toute la richesse d'un corpus tel que Littéracie Avancée pour une approche différente du travail des habiletés rédactionnelles. Je rêverais d'outils qui pourraient combiner la puissance des plates-formes de « e-learning » avec une exploration outillée du corpus telle que nous l'avons menée pour la comparaison entre natifs et non-natifs. Il faudrait permettre aux étudiants de mieux appréhender la dimension textuelle et sa dynamique, assez difficiles à approcher avec des plates-formes basées sur QCM ou textes à trous. Ceux-ci ont leurs mérites, mais ils restent limités : de réels besoins de formation à l'écrit existent, en particulier pour la production d'écrits longs tels que mémoires ou même thèses.

En attendant cet outil idéal, la dernière partie de ce chapitre prolongera ces réflexions par l'exposé d'un projet qui vient d'obtenir un financement de l'ANR[125]. Il ne s'agit donc plus de synthèse de l'existant, mais d'ouvrir sur l'avenir proche en mettant en évidence la façon dont se verra amplifié le champ des recherches conduites jusqu'ici.

6.2 Le projet E-CALM (Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques)

Le titre du projet est, je l'espère, assez explicite : il s'agit, de façon ambitieuse, de connecter et compléter des corpus constitués dans plusieurs laboratoires de recherche, pour mettre à la disposition des chercheurs un vaste ensemble de textes « de la maternelle à l'université » (en fait plutôt du primaire à l'université). Ce corpus donnera lieu à des analyses menées par les

125 Projet ANR-17-CE28-0004-02

différents partenaires du projet, l'ensemble devrait permettre de formuler diverses propositions didactiques.

Le projet se donne trois objectifs principaux :

- structurer et mettre à disposition de la communauté scientifique un vaste corpus d'écrits d'élèves et d'étudiants permettant des analyses quantitatives et des traitements automatiques ;
- caractériser ces écrits et les attentes des enseignants du point de vue de l'acquisition de l'orthographe et de la cohérence, dans des analyses sociologiquement contextualisées ;
- étudier les modalités d'écriture dans les avant-textes (plans, notes, brouillons) et les textes, notamment à travers l'influence réciproque des écrits remis et des interventions des enseignants sur les copies.

Le dernier objectif s'inscrit dans une approche de génétique textuelle qui n'est pas développée au Lidilem, je vais donc surtout évoquer les deux premiers, en me concentrant sur le versant linguistique du projet – je n'exposerai pas en détail le versant sociologique que le poids toujours important des origines sociales sur les compétences scolaires, qui montre l'impossibilité de l'école à contrebalancer les inégalités sociales et culturelles des élèves, nous a conduit à envisager.

6.2.1 Objectif « vaste corpus »

Le projet s'origine dans un double constat : d'une part, les enquêtes PISA sur les compétences des élèves mettent en évidence certaines difficultés avec l'écrit, d'autre part, la production écrite est encore trop peu travaillée dans la scolarité obligatoire (Goigoux, 2016), alors que la capacité à traiter et produire de l'écrit est de plus en plus sollicitée dans la société. On manque en outre de données exploitables et disponibles aux différents paliers de la scolarité obligatoire et au-delà sur les performances des élèves et étudiants en matière de production écrite.

Le projet vise donc en premier lieu la création et la mise à disposition d'un corpus de productions écrites qui pourra précisément fournir des données pertinentes. En rassemblant plusieurs laboratoires qui ont déjà chacun élaboré des recueils de textes, l'objectif est de parvenir à un formatage commun des données pour les harmoniser en un seul ensemble cohérent. Certains corpus concernent l'école primaire, d'autres le collège et lycée, le niveau universitaire est couvert avec Littéracie Avancée (section 6.1.1). Un des premiers chantiers du projet consistera à définir le format commun, et à procéder à de nouveaux recueils afin d'offrir un échantillon de productions textuelles suffisamment varié et propre à permettre d'observer les évolutions des compétences textuelles au long de la scolarité.

En effet, un autre manque repéré dans l'existant réside dans le fait qu'au fur et à mesure de la scolarité, les genres travaillés évoluent, certains types d'écrits n'étant plus guère suscités à l'université, par exemple les textes narratifs, et pas encore travaillés – ou de façon très embryonnaire – au primaire, par exemple les textes argumentatifs ou les synthèses de documents. De ce fait, les comparaisons des productions entre les niveaux scolaires sont rendues difficiles, voire interdites : comment tirer des conclusions sur le développement des habiletés langagières si à un âge l'élève complète une trame narrative et cinq ans plus tard il/elle produit un commentaire de texte ? Nous nous donnerons pour tâche de susciter des textes de divers types pour permettre ces comparaisons.

Dans le même temps, les métadonnées, dont j'ai précédemment montré un exemple à travers celles qui ont été retenues pour Littéracie Avancée, seront aussi harmonisées et complétées autant que possible pour donner, en même temps que les textes, le plus de renseignements possible sur leurs conditions de production. Outre les indications déjà mentionnées, sur le niveau scolaire, la consigne d'écriture, etc. (voir 6.1.1.2), il sera précisé pour chaque texte s'il est produit dans le cadre habituel du cursus scolaire, qu'il soit destiné à une évaluation ou qu'il constitue un entrainement, ou s'il a été suscité par la recherche – puisque nous collecterons de nouveaux textes expressément pour le corpus.

On envisage ainsi un corpus qui sera sans commune mesure avec les corpus scolaires ordinairement travaillés : Elalouf (2011) rend compte de quatre thèses fondées sur des analyses de corpus qui sont de l'ordre d'une centaine de textes, quand Littéracie Avancée comporte à lui seul 330 textes et avoisine le million de mots. Sans céder à la tentation du « gros c'est beau », dénoncée par Péry-Woodley (1995), nous sommes guidés pour la construction de ce corpus par l'hypothèse que la variété des données autorisera des travaux de plus grande ampleur sur les caractéristiques des productions écrites scolaires, selon divers axes d'analyse. Le projet en développe plusieurs, dont j'ai donné un aperçu dans le préambule de cette section, je présente ici les deux axes qui se situent dans la continuité de mes travaux.

6.2.2 « Zoom » sur la cohérence textuelle

Charolles (1978) pointait la difficulté rencontrée par les enseignants, en particulier au primaire, pour faire progresser les élèves au niveau textuel, difficulté qui tient au caractère un peu insaisissable de la cohérence. Qu'est-ce qui fait qu'un texte est cohérent, pourquoi tel ou tel texte d'élève est-il jugé non conforme aux attentes normées en matière de cohérence ? Nombre d'enseignants peinaient à identifier, et encore plus à formuler, des critères précis, applicables aux textes d'élèves. Vingt ans plus tard, une enquête menée par Rondelli (2010) rend compte d'une évolution. À la question « Qu'est-ce que la cohérence textuelle ? », les réponses des enseignants interrogés livrent d'une part une référence récurrente au travail de Charolles cité plus haut, d'autre part une conception de la cohérence comme propriété interne du texte, reposant sur sa logique, son sens, son usage approprié des marques linguistiques du liage et de la continuité (anaphores et substituts notamment). Les travaux de linguistique, en particulier ceux de M. Charolles et B. Combettes, ont irrigué la formation des enseignants, notamment à travers certains ouvrages de préparation au concours[126], qui reprennent explicitement les quatre méta-règles formulées par (Charolles, 1978). Toutefois, remarque Rondelli (2010), la chaine de transposition didactique n'est pas complète. Les enseignants enquêtés envisagent la cohérence comme un phénomène très « surfacique » et très local. Les remarques qu'ils apposent sur les textes témoignent d'un intérêt important accordé aux marques qui construisent des liens interphrastiques, au détriment d'une appréhension globale de la construction du texte et du « monde » qu'il met en place, donc sans percevoir (au moins dans leurs déclarations et dans leurs évaluations en actes) que ce monde peut servir d'interprétant aux divers temps de la narration. Je reprendrai de (Rondelli, 2010 : 74) un exemple que je trouve frappant :

(138) – « Où est le requin ? dit un spécialiste.– Il est là ! crient les enfants en chœur.– On s'en occupe ! dirent les autres spécialistes.– Pouvez-vous nous ramener ? dirent-ils tous en chœur. »

Cet extrait a donné lieu à un jugement d'ambigüité et à des demandes d'explicitation du pronom *ils* (*dirent-**ils** tous en chœur*). De fait, si l'on se place dans une perspective de stricte

126 Notamment, dit Rondelli (2010), celui édité par Hatier, qui compte des linguistes parmi ses auteurs.

continuité textuelle, *ils* devrait avoir comme référent le groupe nominal pluriel le plus proche, donc *les spécialistes*. Mais si l'on considère que la dimension sémantico-pragmatique prend le pas sur la logique interphrastique, aussi bien le discours tenu (une demande d'être ramenés) que la logique de l'échange conversationnel (l'alternance des tours de parole) ou que la répétition de la façon dont la parole est prise (*en chœur* / *tous en chœur*) annulent l'ambigüité potentielle. Le commentaire enseignant en l'occurrence révèle une attention au fonctionnement du pronom « du point de vue de la continuité interphrastique plutôt que de la continuité textuelle » (Rondelli, 2010 : 77).

Le travail que l'on se propose de mener dans le projet E-CALM[127] vise la mise en place de méthodes, outils et ressources pour l'analyse des facteurs de cohérence dans les écrits scolaires. Il reposera sur l'annotation de l'organisation discursive, dans certains textes du corpus, selon trois approches complémentaires qui devraient permettre de saisir l'ensemble des phénomènes textuels à prendre en compte, i. une approche ascendante, ii. une approche descendante, iii. une approche surfacique :

i. à partir d'une segmentation des textes en unités élémentaires seront annotées les relations de cohérence entre les segments – selon la démarche mise en œuvre dans le projet ANNODIS (Ho-Dac & Péry-Woodley, 2014) ;
ii. à partir de schémas macro-structuraux, tels que par exemple le schéma narratif, seront dégagées des unités minimales ;
iii. à partir de marques de surface telles qu'anaphoriques, connecteurs, etc., seront annotés les indices de cohésion textuelle.

On mènera ce travail sur des textes à divers niveaux de la scolarité et surtout dans divers genres textuels, de façon à saisir les « ingrédients » de la cohérence textuelle aussi bien sur des textes narratifs, qui sont en général la première cible des études de la cohérence textuelle et sont surtout la première entrée en écriture au niveau scolaire, que sur des textes argumentatifs, moins étudiés et élucidés. Il sera alors possible de mener des comparaisons entre des genres textuels différents, afin de mettre au jour des différences ou des convergences au niveau de ce qui fait la cohérence, et de cerner l'évolution de la maitrise de la cohérence au fil de la scolarité.

Le travail que j'ai mené sur les intertitres pourra trouver là un prolongement naturel : on a vu (chapitre 4) qu'ils sont volontiers porteurs d'indices de cohésion et que par ailleurs, la RST Rhetorical Structure Theory), un modèle de relations de cohérence, leur assigne toujours la même relation *Preparation* par rapport au texte. L'annotation projetée permettra de poursuivre et d'affiner la réflexion sur les relations que les intertitres entretiennent avec le texte.

Les textes annotés constitueront une ressource « grandeur nature » pour la formation des enseignants, aussi bien au primaire qu'au secondaire, qui permettra d'une part la prise de conscience du fait que la cohérence textuelle ne se réduit pas à un bon usage des liaisons interphrastiques, d'autre part un travail sur le geste d'évaluation et de correction de textes d'élèves, geste lui aussi très entaché de stéréotypes et révélateur des normes implicites mises en œuvre par les enseignants dans la lecture des textes d'élèves (Pilorgé, 2010).

6.2.3 « Zoom » sur les normes enseignantes

Hormis les travaux que j'ai cités de F. Rondelli et J.-L. Pilorgé, les commentaires et indications apportés par les enseignants sur les travaux d'élèves ou étudiants n'ont guère été étudiés.

127 Je cite et reformule dans toute cette fin de section une partie du texte du projet co-rédigé par les divers partenaires.

Pourtant, aussi bien les passages textuels qui font réagir les enseignants que le contenu des commentaires qu'ils inscrivent sont riches d'enseignement sur leur conception du texte et sur leurs attentes, et sont révélateurs du « contrat » didactique plus ou moins explicite qui guide leur lecture.

Pilorgé dégage de l'analyse de ces interventions écrites des enseignants « cinq postures caractéristiques de la variété des attitudes adoptées » (2010 : 93), plusieurs postures pouvant être adoptées successivement par un même correcteur :

- le « gardien du code » s'en tient à la correction formelle du texte et ne s'implique guère en tant que lecteur ;

- le « lecteur naïf » lit le texte en référence au monde réel, sans prendre en compte la construction du monde du texte opérée par l'élève-auteur ;

- le « stimulus-réponse » examine la conformité de la production de l'élève au stimulus que constitue la consigne d'écriture ;

- l'« éditeur » prend en compte le projet de l'élève, entre dans le monde du texte et formule des indications de nature à permettre une amélioration du texte cohérente avec l'intention de l'élève ;

- le « critique » traite le texte comme une production littéraire, propose des appréciations sur un plan esthétique, entre en dialogue avec l'auteur-élève.

Pilorgé (2010) note que plus le texte est jugé éloigné de la norme textuelle, plus le correcteur se centre sur ses aspects formels au détriment d'une lecture littéraire qui prendrait en compte le projet de l'auteur-élève et aiderait ce dernier à produire des réécritures propices à la réalisation de ce projet. Finalement, Rondelli (2010) comme Pilorgé (2010) pointent

> [l]a tendance de l'univers scolaire à instituer un cadre normatif parfois beaucoup plus contraignant que celui que mettent en œuvre les pratiques sociales de référence, à construire des hypernormes (Pilorgé, 2010 : 100).

Notre propos est de nous appuyer sur des interventions spontanées des enseignants qui nous ont fourni les textes du corpus ou sur des corrections suscitées dans le cadre du projet, pour élaborer une typologie de ces interventions qui mette en évidence les normes sous-jacentes mobilisées. Il s'agira, en prenant en compte la modalité de l'intervention (du simple soulignement jusqu'à la proposition de réécriture en passant par des commentaires plus ou moins étayés) et l'objet ou phénomène sur lequel porte l'intervention (orthographe, marques de cohésion, conformité au genre, etc.), de dégager des types et de les corréler aussi bien au genre textuel qu'au niveau scolaire.

Pour donner un peu de chair à cette partie du projet, je trace maintenant les grandes lignes d'un travail qui est en cours sur les interventions sur des textes du corpus Littéracie Avancée.

Parmi les 339 textes actuellement réunis, 81 comportent des interventions apposées par les enseignants, soit sous la forme de surlignement, soit sous la forme de commentaire ajouté au texte avec la fonction « commentaire » du traitement de texte. Ces interventions sont destinées aux auteurs-étudiants en vue d'une réécriture, dans la situation d'allers-retours d'un écrit, situation classique au niveau universitaire. Une première étude exploratoire[128] a porté sur 30 textes, qui représentent plus de 760 commentaires. La typologie ébauchée et que je présente ci-

128 Que j'ai menée avec le concours d'Émilie Charles, ancienne doctorante du Lidilem, que je remercie de son assistance.

dessous est une première tentative de rendre compte du niveau sur lequel porte l'intervention, de son étendue (simple signalement ou proposition de réécriture) et de l'intention de l'enseignant-correcteur. On verra avec les exemples que l'enseignant du supérieur entre dans un véritable dialogue avec l'étudiant, par texte interposé, pour aider celui-ci à construire aussi bien les normes et attentes du genre, qu'une réflexion disciplinaire et scientifique. De façon schématique, les interventions se situent sur quatre dimensions qui ne sont pas étanches.

1. Le niveau formel

Les commentaires pointent des défaillances formelles, que ce soit sur le plan orthographique (qui inclut la ponctuation), sur celui de la mise en forme ou sur les choix de formulation qui influent sur la clarté et l'intelligibilité du propos. Dans les exemples, j'indique en italique le texte cible du commentaire et celui-ci en caractères droits.

Orthographe

(139) *était*
été [Littéracie avancée]

(140) *chercheurs*
11 erreurs d'orthographe sur 20 lignes, c'est beaucoup trop: tout votre mémoire est à peu près sur le même modèle: il faut faire une correction intégrale sinon vous allez être très pénalisée par l'orthographe. [Littéracie avancée]

Mise en forme

(141) *« les professeurs professionnels ne se sentent pas responsables du niveau d'acquisition langagière des élèves tout comme les professeurs de français ne se sentent pas responsables de ce que les élèves ne savent pas faire en atelier »*
Choisissez l'italique ou les guillemets, mais pas les deux [Littéracie avancée]

(142) *2007:141*
il faudrait harmoniser les espaces avant et après les deux points dans la parenthèse indiquant l'année et la page. On trouve de tout chez vous! [Littéracie avancée]

Formulation, clarté, intelligibilité

(143) *relationnel*
pour moi c'est trop oral [Littéracie avancée]

(144) *d'élaborer l'écriture*
Pourquoi élaborer l'écriture? pourquoi pas «écrire» tout simplement? soyez plus claire et plus simple dans vos tournures [Littéracie avancée]

Certains commentaires sont à mi-chemin entre la dimension formelle et la dimension conceptuelle, témoignant ainsi du fait que le mot aide à penser :

(145) *Le taux de difficulté/facilité perçue :*
Je parlerais plutôt de «degré de réussite», qui me semble plus clair [Littéracie avancée]

2. Le niveau conceptuel

Il s'agit ici moins de travailler la forme que ce que la formulation révèle de la conceptualisation qui s'élabore. La frontière avec la catégorie précédente est assez floue dans la mesure où la conceptualisation est saisie à travers les formulations choisies, cependant, dans ses

commentaires, l'enseignant-e ne cible pas en premier lieu la forme mais s'attache au contenu. Deux mouvements sont privilégiés dans les commentaires : la rectification, qui vise à corriger une conception qui semble erronée, ou l'incitation au développement de la pensée, à l'approfondissement dans la direction que le propos de l'étudiant suggère ou qui est indiqué par le commentaire.

Rectification

(146) *Nous pouvons imaginer qu'au CM2, la motivation principale de ces élèves est celle liée aux notes.*
Non, chez l'élève (même plus grand), la motivation est le plus souvent liée à l'intérêt qu'il prend à étudier, même si la matière n'est pas notée. D'où l'intérêt de choisir avant tout des dispositifs qui stimulent l'intérêt de l'élève, c'est-à-dire des dispositifs qui font appel à son intelligence. [Littéracie avancée]

(147) *Ces derniers expliquent ainsi les volontés marquées des professeurs de réformer l'orthographe*
Info à nuancer avec les travaux actuels, voir par ex. le n° 19 de la revue glottopol (http://glottopol.univ-rouen.fr/numero_19.html) [Littéracie avancée]

Cet exemple (147) montre une stratégie qui, là aussi, oscille entre la rectification – qui consiste à indiquer à l'étudiant une erreur ou une approximation conceptuelle – et la suggestion d'approfondissement. Il ne nie pas frontalement l'assertion de l'étudiant mais renvoie celui-ci à des références susceptibles de compléter ses connaissances et de l'amener à nuancer l'affirmation qui est commentée. La différence avec le type suivant est ténue et tient au fait qu'il y a réfutation du propos de l'étudiant.

Approfondissement / développement

(148) *En fonction des différents profils d'élèves, ces activités se sont révélées plus ou moins efficaces. En effet, les élèves sans difficultés particulières se sont montrés réceptifs dans la plupart des cas. En revanche, en ce qui concerne les élèves les plus*
ce paragraphe est le seul retour que vous faites sur ce que ce stage et ce mémoire a pu vous apprendre pour votre pratique professionnelle concernant la compréhension en lecture. Je pense que c'est là qu'il faudrait creuser davantage. [Littéracie avancée]

(149) *sur trois documents*
OK pour la synthèse, mais attention: pour le mémoire, il faudrait trouver en complément des documents plus récents de champ très labouré [Littéracie avancée]

Dans (149), sont simultanément indiquées deux pistes que l'étudiant aura à explorer ultérieurement : accroitre quantitativement ses références (la phrase entière est « notre synthèse prendra appui sur trois documents » et l'on voit que c'est ce nombre qui donne lieu à commentaire) et chercher des documents plus récents. Cette demande est justifiée par une attente générique et disciplinaire : le travail demandé, une synthèse, tolère un petit nombre de références, en revanche, le travail à venir, le mémoire, requerra des références plus étoffées. Le commentaire trace en filigrane le cadrage inhérent au genre : on ne fait pas un mémoire de recherche à partir de si peu de références quand il en existe pléthore (« champ très labouré ») et plus récentes.

D'autres commentaires renvoient au niveau textuel, intriqué dans la structure « programmée » par le genre.

3. Le niveau textuel et générique

Sont rassemblées sous cette bannière les interventions qui portent sur la construction du texte et les stratégies d'organisation et d'exposition qu'elle requiert, et celles qui rappellent les normes éditoriales du genre, notamment en matière de citation et bibliographie.

Stratégie / construction argumentative / textuelle

(150) *c'est par lui que le spectateur va enfin pouvoir comprendre le message, que nous verrons dans notre deuxième partie, créé par l'image*
A mettre dans la seconde partie alors? [Littéracie avancée]

(151) *Annexes.*
si vous mettez tous ces documents en annexe, il faut y renvoyer lors de votre rapport, sinon il vaut mieux en supprimer quelques unes. N'oubliez pas de les numéroter et de faire des renvois précis (cf. annexe 1) par ex./ [Littéracie avancée]

Norme éditoriale

(152) *Introduction*
dans ce type de dossier, l'introduction est un seul paragraphe (pas d retour à la ligne). (sic) [Littéracie avancée]

(153) *« les documents utilisés par les formatrices sont éloignés des documents susceptibles d'être manipulés par les*
Attention : pour chaque citation, indiquer l'année et la page par ex: «(Adami, 2008, p. 34)» [Littéracie avancée]

Ces divers exemples montrent que la distinction entre les deux types tient toute entière dans un recours à la norme, marqué par des formulations assertives voire impératives *vs* des formulations plus ouvertes, qui laissent le choix à l'auteur pour la construction de son texte ou son argumentation.

La dernière dimension sollicitée par les commentaires est liée à la discipline.

4. La dimension disciplinaire

L'écrit produit par l'étudiant, qu'il s'agisse d'un dossier, d'un rapport de stage, d'une synthèse d'articles de recherche, d'un mémoire, est (aussi ? surtout ? principalement ?) un lieu d'acculturation aux concepts et méthodes de la discipline. Les concepts sont travaillés via la rectification ou la suggestion de développement ou d'approfondissement – extraits (146) à (149) –, nous en avons distingué les interventions qui suivent, qui ciblent les problèmes méthodologiques.

Méthodologie

(154) *Le site multimania.fr*
Ici aussi, vous pouvez vous référer à des ouvrages plus spécialisés que le site multimania (par ex une grammaire telle que celle de Riegel, M., Pellat, J.C., & Rioul, R. (2004). Grammaire méthodique du français. Paris : Presses universitaires de France.) [Littéracie avancée]

(155) *du verbe «arriver», tout comme précédemment avec le verbe «venir*
Attention: c'est le même verbe en anglais. Il n'y a pas de raison de commenter votre traduction. L'analyse que vous faites doit porter sur le corpus original. Donc à traiter avec les autres cas de polysémie. [Littéracie avancée]

Ce qui est pointé avec ces commentaires, c'est la manière de s'y prendre de l'étudiant, non conforme à la démarche d'analyse ou de construction du savoir attendue par la discipline. On voit avec (155) le retentissement d'une erreur méthodologique (ne pas analyser le corpus original mais sa traduction) sur la construction textuelle. Au bout du compte, le respect de cette démarche garantit la recevabilité du texte à tous les niveaux.

Tous les exemples qui précèdent montrent des interventions ciblées sur des passages jugés défaillants de la production. Un petit nombre manifeste l'approbation et l'encouragement du lecteur-correcteur.

Approbation

(156) *commencé à rédiger un carnet de bord relatant ce que j'avais prévus de faire lors de la séance, ce que j'arrivais à faire ainsi que la réaction des élèves face au dispositif. Je le remplissais directement après la séance. Effectivement, une fois ma séance*
Bien: démarche rigoureuse [Littéracie avancée]

(157) *hypothèses*
Une très bonne entrée en matière, à la fois claire et précise. [Littéracie avancée]

De nombreux commentaires approbatifs se résument à « oui », « ok » ou « bien ».

La typologie que je viens de présenter est encore préliminaire, elle mérite assurément d'être retravaillée, voire totalement révisée. Elle est de toute façon amenée à évoluer dans le cadre du projet E-CALM. Elle est en effet bâtie à partir des écrits d'étudiants et des catégories mobilisées par les enseignants pour guider les étudiants. Il sera sans doute fructueux de se distancier de ces écrits et de ces catégories au profit de types qui seraient appropriés aussi aux interventions sur les textes d'élèves, lesquelles mobilisent sans doute d'autres catégories – mais c'est précisément ce que la recherche se propose de vérifier.

Après la typologie des interventions des enseignants, il est envisagé d'utiliser l'ensemble de ces résultats pour la formation des enseignants du primaire et du secondaire au geste de correction, d'abord par une prise de conscience des objets et des niveaux privilégiés par les interventions, ensuite par des travaux spécifiques autour de l'évaluation. En prenant appui sur un corpus restreint constitué à cet effet au cours de l'étude de la cohérence, l'idée est de travailler avec les enseignants à une comparaison des critères sur lesquels s'appuie leur évaluation et des annotations produites par l'équipe de chercheurs. L'objectif est de provoquer une décentration des enseignants pour une évaluation plus soucieuse de toutes les caractéristiques des textes.

6.3 Synthèse-bilan du chapitre

Ce dernier chapitre donne corps à une recherche orientée vers l'enseignement de la langue, au sens large. Pour prendre en compte les besoins du public destinataire de l'enseignement, a été constitué et est diffusé un corpus d'écrits universitaires. Ce corpus est enrichi d'un balisage xml conforme aux préconisations de la TEI, dans lequel sont repérées les structures textuelles (découpage en sections, en paragraphes, titres, etc.) et sont insérées des métadonnées qui font de ce corpus une ressource exploitable au-delà des chercheurs qui l'ont constitué.

Deux exploitations de ce corpus sont actuellement en cours :

- dans une optique de description des compétences textuelles et de comparaison de ces compétences avec celles d'un public allophone, on s'est intéressé à une structure

typique de l’écrit scientifique « comme ... Verbe », ce qui a montré un certain nombre de difficultés de maitrise de cette structure partagées par tous les étudiants, natifs et non-natifs ;

- dans une optique de formation des étudiants, un cours en-ligne a été élaboré, dont la particularité réside dans le fait qu’il s’appuie sur les erreurs et réussites des étudiants pour proposer à ce même public des exercices destinés à susciter une posture réflexive à l’égard de la langue et du texte.

Ce corpus sera à terme intégré, totalement ou en partie, dans un corpus plus vaste, couvrant tous les niveaux scolaires, du primaire au M2, et ce dans le cadre du projet ANR E-CALM débutant en janvier 2018.

Les recherches sur ce grand corpus couvriront diverses problématiques de la textualité et de l’enseignement de la production de textes (orthographe, cohérence textuelle, génétique du texte, correction des enseignants, séquences didactiques utilisant le corpus…). J’ai focalisé une partie du chapitre sur les parties du projet qui concernent :

- la cohérence textuelle, en mettant en avant les problèmes que celle-ci continue à poser aux enseignants du primaire et du secondaire ;
- les interventions de correction des enseignants sur les productions écrites, et j’ai là donné un aperçu d’une étude préliminaire menée sur les commentaires portés par des enseignants du supérieur sur les écrits étudiants.

Chapitre 7 - Conclusion et perspectives

J'ai tout au long de cette synthèse tenté de montrer l'articulation entre différents travaux dont l'un des dénominateurs communs est de s'ancrer dans le texte, en prenant au sérieux sa dynamique, qui a des effets sur le lexique et sur les objets que l'architecture textuelle contribue à créer. L'ensemble des travaux menés s'appuie sur un outillage linguistique qui mobilise lorsque nécessaire des concepts et des techniques de TAL.

L'essentiel de mon travail de recherche s'est orienté vers l'analyse et le traitement de textes spécialisés, se focalisant plus particulièrement, dans la période récente, sur l'écrit scientifique et académique. Pour ces études, j'ai participé à la constitution de corpus dont l'un, Littéracie Avancée, est libre de droits et disponible pour la recherche. Il permet un travail sur deux versants : la description des compétences rédactionnelles et la réflexion sur l'enseignement.

Deux thèses en cours que je co-encadre, celle de Cindy De Amaral (commencée en 2016, dirigée par C. Frier) et celle de Luca Pallanti (commencée en 2017, dirigée par C. Brissaud), combinent de même un recueil de production d'apprenants (lycéens en lycée professionnel pour la première, étudiants en IUT pour le second), une description en terme de compétences textuelles et une expérimentation didactique. Elles feront donc progresser la recherche dans ces directions tout en assurant la formation à la recherche de leurs auteurs.

7.1 La formation des scientifiques du langage

Ma motivation première pour l'obtention d'une habilitation à diriger des recherches est de … diriger des recherches, c'est-à-dire des thèses. Après cette tautologie en forme de boutade, je voudrais intégrer dans cette conclusion un exposé de ce qui me parait important et même essentiel pour former un-e futur-e « scientifique du langage » et que je projette donc de mettre en œuvre dans mon rôle de formation à la recherche. Je ne reviendrai pas sur les incontournables de la discipline, faire un état de l'art, construire une problématique, etc., mais me focaliserai sur ce que le virage vers une linguistique de corpus implique à mon sens d'ajouter dans la discipline.

1. Porter attention aux données

Les sciences du langage sont empiriques. En dépit de la connotation parfois négative du terme, c'est une caractéristique qu'il faut assumer à travers une attention aux données selon deux points de vue : les propositions et descriptions doivent être fondées sur des données, dont elles doivent rendre compte sans multiplier les exceptions, et elles doivent être adéquates aux données. Ce qui signifie de faire la part des faits et de leur interprétation, en dissociant les faits de l'analyse qui en est produite. Je compte engager les doctorants avec lesquels je travaillerai à mettre en place les procédures adaptées pour garder trace des données initiales sur lesquelles ils travaillent et de façon distincte des analyses qui en sont produites, de façon à toujours pouvoir retrouver le chemin qui les a menés à leurs conclusions, à pouvoir réexaminer les mêmes données avec un œil nouveau, ou à pouvoir échanger ces données avec d'autres chercheurs.

Une bonne pratique, je l'ai indiqué au chapitre 2, consiste, quand l'analyse fait intervenir des jugements, à confronter des jugements de chercheurs différents et à mesurer leur taux d'accord. C'est ce que nous (l'équipe du projet Termith) avons mis en place pour la thèse de S. Hatier

(2016), en confiant pour analyse les mêmes données à plusieurs chercheurs du projet, et c'est à mon sens le meilleur moyen de garantir la validité des conclusions.

2. Se doter d'outils et de méthodologie solides

Le travail sur les données ne peut être convenablement mené qu'avec les outils adéquats. Concordanciers, logiciels d'exploration, d'annotation, de traitements textométriques et lexicométriques peuvent désormais faire partie de la panoplie du chercheur en sciences du langage (voire même en sciences humaines) comme le microscope fait partie de la panoplie du biologiste. Leur maitrise n'est pas seulement une maitrise technique, elle passe par une réflexion méthodologique qui se construit au fur et à mesure d'une thèse. Le jeune chercheur immanquablement s'interroge sur les outils et procédures à mettre en œuvre au fur et à mesure de ses explorations des données, sur les places respectives des approches quantitatives et qualitatives (qui ne sont au bout du compte que les deux faces d'une même science, d'après Laflamme, 2007), sur les conclusions qui peuvent être validement tirées…

Le travail sur corpus donne une place privilégiée à cette réflexion méthodologique, seule à même de garantir que la trituration des données par logiciel interposé ne soit pas une fin en soi. Il est facile pour un jeune chercheur de céder à la fascination des données, il ne faut pas oublier que leur travail s'inscrit dans une problématique de recherche qui seule donne sens aux explorations, quelles qu'elles soient.

3. Construire un texte intelligible

L'exigence peut paraitre surprenante voire même superfétatoire. Bien sûr, pensera-t-on, qu'un doctorant doit rendre un texte intelligible et construit ! Bien sûr qu'un chercheur doit produire des textes de recherche intelligibles… Mais passé ce premier mouvement de surprise face à une évidence, remémorons-nous les thèses avec lesquelles il a fallu lutter pour en tirer compréhension, les articles que l'on finit de lire en se disant que l'on ne comprend pas bien où « ça » veut en venir.

Même si la production scientifique n'est pas de la production littéraire, il est du devoir d'un chercheur de prendre en compte ses lecteurs, ce qui signifie concrètement assurer leur guidage, rendre explicites les relations au sein du texte, rendre visible son architecture en tant qu'elle est un reflet de la construction de la connaissance que le texte opère.

Nombre de recherches que j'ai citées dans cette synthèse reposent sur le postulat qu'« à travers le travail sur le texte, c'est son intelligibilité qui se travaille, et le raisonnement qui se construit »[129]. L'attention aux formulations et à la construction textuelle dans une thèse ou un article de recherche n'est pas seulement une attention portée au lecteur, elle est un outil de l'élaboration du raisonnement et de la clarification cognitive des concepts. Engager le jeune chercheur dans un travail sur la forme, c'est aussi l'aider à élaborer le contenu, ce qui est à mon sens le rôle premier du directeur de thèse.

J'en viens donc aux perspectives de recherche qui se dessinent sur trois axes.

129 Extrait d'une présentation donnée en 2015 en collaboration avec Fanny Rinck lors de la journée « Corpus et didactique » organisée annuellement à Valence par C. Cavalla et L. Hartwell.

7.2 Perspectives de recherche

1. Penser une linguistique outillée

L'essor de la linguistique de corpus en France ces dix dernières années déplace quelque peu le champ des questions originelles. Le passage de l'introspection à une linguistique qui se fonde sur des données authentiques avait mis au premier plan la problématique de la constitution de ces corpus. Même si les questions relatives à la représentativité des données, aux choix de constitution, aux biais éventuels, ne sont pas totalement résolues, elles ont perdu de leur acuité dans le sens où désormais, me semble-t-il, les recherches en linguistique ne cèdent plus à la naïveté de croire que la vérité est dans le corpus et qu'il suffit de « regarder dans les corpus » pour atteindre le fonctionnement de la langue. S'impose l'idée (ou même la constatation) qu'un corpus est un construit, un objet, qui de ce fait ouvre certaines analyses et en ferme d'autres. Les nouvelles questions qui surgissent alors doivent à mon sens porter sur la méthodologie : que fait-on des corpus, quelles données en extrait-on, quels observables construit-on, comment, quelle transparence assure-t-on par rapport à la communauté des chercheurs, quelle reproductibilité des résultats ? Seule une claire conscience de ce que l'on fait au bout du compte, c'est-à-dire des démarches d'analyse et des conséquences des choix opérés à chacune des étapes des recherches, peut faire progresser la linguistique dite « de corpus ».

Le travail de la matière « corpus » implique un outillage, que ce soit pour l'exploration des données, pour leur enregistrement, pour leur analyse. Cet outillage ne doit pas oblitérer la réflexion linguistique, il doit être l'instrument qui permet d'atteindre les phénomènes, pas le guide de la réflexion. Il me semble donc nécessaire de poursuivre des travaux qui pensent les outils, dans tous les sens d'une telle expression, c'est-à-dire qui œuvrent à la définition d'usages des outils disponibles, aux combinaisons de techniques de TAL, de concordanciers, de mesures statistiques pour produire des analyses, mais aussi qui dévoilent les implicites de l'outillage, les biais éventuels comme les acquis, dans une réflexion plus épistémologique.

La formation des jeunes linguistes doit permettre de réaliser en un seul individu l'ingénieur qui maitrise les techniques et peut s'impliquer dans la conception des outils et le chercheur qui élabore les problématiques et produit des analyses.

2. Connecter la linguistique et d'autres disciplines

Il ressort de ce qui précède une connexion évidente avec l'informatique à travers le traitement automatique des langues. Faire progresser tous les traitements, étiquetages et annotations automatiques, balisages, repérages de structures et de phénomènes, qui enrichissent des données brutes et augmentent le potentiel d'analyse est une perspective à inscrire dans le champ de la linguistique de corpus.

La prise en compte du niveau textuel est encore trop absente des traitements automatiques mais aussi des études sur corpus. Pourtant le texte est le support d'opérations cognitives différenciées (comme le montrent les travaux cités dans le chapitre 4). La production de travaux propres à alimenter la construction d'expérimentations psycholinguistiques, qui en retour permettront d'amender les descriptions des phénomènes, par exemple sur la saillance cognitive, sur la facilitation de la compréhension de certaines structures ou organisations textuelles, sur les processus impliqués dans la production écrite, est encore un chantier ouvert. Nombre de travaux linguistiques, ceux que je mène sur les intertitres en font partie, postulent des traitements cognitifs à la source ou comme conséquence de structures linguistiques observées. Quelles vérifications sont possibles ? Comment les mettre en œuvre ? Quels biais

éviter ? De la même manière que certaines études de psychologie cognitive ont pris conscience de la nécessité d'inclure un savoir linguistique dans la définition des expérimentations (Lemarié, Lorch & Péry-Woodley, 2012), les études linguistiques qui débouchent sur des hypothèses cognitives doivent se soucier de la plausibilité psychologique de leurs conclusions.

Pour ma part, je laisse ce chantier en friche dans l'immédiat, étant actuellement mobilisée par d'autres projets, mais il fait partie de mes perspectives d'avenir, à court ou moyen terme.

3. Connecter la linguistique et la société

J'ai dans la dernière partie de cette synthèse défendu l'idée d'une implication de la recherche « dans la cité ». Il est de nombreux terrains sur lesquels les linguistes peuvent utilement faire fructifier leur expertise, pour le bénéfice de la société comme pour celui de la recherche. Celui sur lequel je m'inscris est l'enseignement, en particulier de la production textuelle. Des travaux linguistiques orientés vers l'enseignement peuvent adopter plusieurs perspectives :

- une réflexion didactique, avec la création de matériel pour l'enseignement, ressources et séquences ;
- des descriptions linguistiques qui aident à cerner les spécificités des objets d'apprentissage, qui mettent en lumière les zones résistantes des textes et de la langue ;
- des descriptions spécifiques des productions des apprenants, en termes d'habiletés et de compétences ;
- des études du développement, au fil du temps et au fil de la scolarité, de ces habiletés et compétences.

Les recherches menées sur les intertitres et globalement sur la structuration textuelle demandent à être mises en relation selon ces perspectives avec l'ensemble des travaux sur la cohérence et la cohésion textuelle, pour situer les observations dans l'économie générale du texte. C'est le chantier qui s'ouvre avec le projet E-CALM, que j'ai présenté dans le chapitre 6. Il représente l'actualité qui s'ouvrira en 2018.

Références

Adam J.-M. (1992). *Les textes, types et prototypes*. Paris : Nathan.

Adam J.-M. (2001). Types de textes ou genres de discours ? Comment classer les textes qui disent de et comment faire ? *Langages 141*, pp. 10-27.

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

Apothéloz D. (1995a). Nominalisations, référents clandestins et anaphores atypiques. *TRANEL 23*, pp. 143-173.

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

Apothéloz D. & Reichler-Béguelin M.-J. (1995). Construction de la référence et stratégies de désignation. *TRANEL 23*, pp. 227-271.

Assadi H. & Bourigault D. (2000). Analyses syntaxique et statistique pour la construction d'ontologies à partir de textes. In J. Charlet, M. Zacklad, G. Kassel, D. Bourigault (Éds), *Ingénierie des connaissances Evolutions récentes et nouveaux défis*, pp. 243-255. Paris : Eyrolles.

Aussenac-Gilles N. & Condamines A. (2000). Entre textes et ontologies formelles : les bases de connaissances terminologiques. In M. Zacklad, M. Grundstein (Éds), *Ingénierie et capitalisation des connaissances*, pp. 153-177. Paris : Hermès.

Aussenac-Gilles N. & Jacques M.-P. (2006). Designing and Evaluating Patterns for Ontology Enrichment from Texts. In *Managing Knowledge in a World of Networks* (Vol. 4248/2006), pp. 158-165. Berlin / Heidelberg : Springer.

Aussenac-Gilles N. & Jacques M.-P. (2008). Designing and evaluating patterns for relation acquisition from texts with Caméléon. *Terminology 14*(1), pp. 45-73.

Aussenac-Gilles N. & Séguéla P. (2001). Les relations sémantiques : du linguistique au formel. *Cahiers de Grammaire 25*, pp. 175-198.

Bastuji J. (1974). Aspects de la néologie sémantique. *Langages 36*, pp. 6-19.

Berrendonner A. (1990). Attracteurs. *Cahiers de linguistique française 11*, pp. 149-158.

Biber D. (1988). *Variations across speech and writing*. Cambridge / New York : Cambridge University Press.

Biber D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge / New York : Cambridge University Press.

Boch F. (2013). Former les doctorants à l'écriture de la thèse en exploitant les études descriptives de l'écrit scientifique. *Linguagem em (Dis)curso 13*(3), pp. 543-568.

Boch F. & Rinck F. (2016). Anaphores démonstratives dans les écrits d'étudiants de Master. Comparaison avec les pratiques expertes. *Linx 72*.

Borillo A. (1996). Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie. *LINX 34-35*, pp. 113-124.

Borillo A. (1997). Statut et mode d'interprétation des noms collectifs. In C. Guimier (Éd.), *Co-texte et calcul du sens Actes de la table ronde tenue à Caen*, pp. 105-121. Caen : Presses Universitaires de Caen.

Bosredon B. (1997). *Les titres de tableaux. Une pragmatique de l'identification*. Paris : PUF.

Boulton A. & Pérez-Paredes P. (2014). ReCALL special issue: Researching uses of corpora for language teaching and learning. *ReCALL 26*(2), pp. 121-127.

Bourigault D. (1994). *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes* (Doctorat Nouveau Régime). École des Hautes Études en Sciences Sociales, Paris.

Bourigault D., Aussenac-Gilles N. & Charlet J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes. *Revue d'intelligence artificielle 18*(1), pp. 87-110.

Bourigault D. & Charlet J. (2000). Ontologie et textes. In *Complément aux Actes de la conférence IC'2000, Journées Francophones d'Ingénierie de la Connaissance*, pp. 7-8. Toulouse : Institut de Recherche en Informatique de Toulouse.

Bourigault D. & Fabre C. (2001). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire 25*, pp. 131-151.

Bourigault D., Fabre C., Frérot C., Jacques M.-P. & Ozdowska S. (2005). Syntex, analyseur syntaxique de corpus. In *TALN'2005* (Vol. 2), pp. 17-20. Dourdan.

Bourigault D. & Slodzian M. (1999). Pour une terminologie textuelle. *Terminologies nouvelles 19*, pp. 29-32.

Boutet J. (2005). Pour une activité réflexive sur la langue. *Le français aujourd'hui 148*(1), pp. 65-74.

Bouveret M. (1997). Le terme : une dénomination au sens réglé. In Équipe de Recherche en Syntaxe et Sémantique (Éd.), , pp. 115-126. *Deuxièmes rencontres Terminologie et Intelligence Artificielle, TIA 97*, Toulouse : Équipe de Recherche en Syntaxe et Sémantique.

Branca-Rosoff S. (1999). Des innovations et des fonctionnements de langue rapportés à des genres. *Langage et société 87*, pp. 115-129.

Busquets J., Vieu L. & Asher N. (2001). La SDRT : une approche de la cohérence du discours dans la tradition de la sémantique dynamique. *Verbum 23*(1), pp. 73-101.

Buyssens E. (1969). La grammaire générative selon Chomsky. *Revue belge de philologie et d'histoire 47*(3), pp. 840-857.

Cadiot P. (1994). Représentations d'objets et sémantique lexicale : Qu'est-ce qu'une boîte ? *French Language Studies 4*, pp. 1-23.

Cappeau P. & Roubaud M.-N. (2005). *Enseigner les outils de la langue avec les productions d'élèves: cycles 2 et 3*. Paris : Bordas.

Carletta J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics 22*(2), pp. 249-254.

Carter-Thomas S. & Jacques M.-P. (2017). Interdisciplinary and interlinguistic perspectives on Academic Discourse: the mode variable. *CHIMERA: Romance Corpora and Linguistic Studies 4*(1), pp. 1-11.

Cellier J.-M. & Terrier P. (2001). Le rôle de la mise en forme matérielle dans le traitement cognitif de consignes. *Langages 141*, pp. 79-91.

Chafe W. (1992). The importance of corpus linguistics to understanding the nature of language. In J. Svartik (Éd.), *Directions in Corpus Linguistics*, pp. 79-97. Berlin / New-York : Mouton de Gruyter.

Chambers A. (2010). L'apprentissage de l'écriture en langue seconde à l'aide d'un corpus spécialisé. *Revue française de linguistique appliquée XV*(2), pp. 9-20.

Charolles M. (1978). Introduction aux problèmes de la cohérence des textes. *Langue Française 38*, pp. 7-41.

Charolles M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques 57*, pp. 3-13.

Charolles M. (1996). Quand intervient le contexte dans la résolution des ambiguïtés. *Scolia 6*, pp. 163-184.

Charolles M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. *Cahier de Recherche Linguistique 6*, pp. 1-73.

Charolles M. (2004). Sinon d'hypothèse négative. In A. Auchlin, M. Burger, L. Filliettaz, A. Grobet, J. Moeschler, L. Perrin, … L. de Saussure (Éds), *Structures et discours. Mélanges offerts à Eddy Roulet*, pp. 167-182. Québec : Nota Bene.

Charolles M. & Combettes B. (1999). Contribution pour une histoire récente de l'analyse du discours. *Langue Française 121*, pp. 76-116.

Charolles M. & Péry-Woodley M.-P. (2005). Introduction au numéro « Les adverbiaux cadratifs ». *Langue française 148*(4), pp. 3-8.

Collet T. (2000). *La réduction des unités terminologiques complexes de type syntagmatique* (Ph. D.). Université de Montréal.

Collet T. (2003). A two-level grammar of the reduction processes of French complex terms in discourse. *Terminology 9*(1), pp. 1-27.

Collet T. (2009). La manière de signifier du terme en discours. *Meta: Journal des traducteurs 54*(2), pp. 279-294.

Combettes B. (1988). *Pour une grammaire textuelle : la progression thématique*. Bruxelles / Paris : De Boeck / Duculot.

Condamines A. (2000). Chez dans un corpus de sciences naturelles Un marqueur de relation méronymique ? *Cahiers de Lexicologie 77*, pp. 165-187.

Condamines A. (2005). Anaphore nominale infidèle et hyperonymie : le rôle du genre textuel. *Revue de Sémantique et Pragmatique* (18), pp. 23-42.

Condamines A. (2006). Avec et l'expression de la méronymie : l'importance du genre textuel. In G. Kleiber, C. Schnedecker, A. Thyssen (Éds), *La relation « Partie - Tout »*, pp. 633-650. Leuven : Peeters.

Condamines A. & Jacques M.-P. (2006). Le repérage de l'hyperonymie par un faisceau d'indices : mise en question de la notion de « marqueur ». In *Journée « Textes et connaissances », Semaine de la Connaissance* (Vol. 1), pp. 185-194. Nantes.

Condamines A. & Narcy-Combes J.-P. (2015). La linguistique appliquée comme science située. In F. Carton, J.-P. Narcy-Combes, M.-F. Narcy-Combes, D. Toffoli (Éds), *Cultures de recherche en linguistique appliquée*. Paris : Riveneuve éditions.

Condamines A. & Rebeyrolle J. (2000). Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In J. Charlet, M. Zacklad, G. Kassel, D. Bourigault (Éds), *Ingénierie des connaissances Evolutions récentes et nouveaux défis*, pp. 225-241. Paris : Eyrolles.

Corbin P. (1980). De la production des données en linguistique introspective. In A.-M. Dessaux-Berthonneau (Éd.), *Théories linguistiques et traditions grammaticales*, pp. 121-177. Lille : Presses Universitaires de Lille.

Cori M. & David S. (2008). Les corpus fondent-ils une nouvelle linguistique ? *Langages 171*(3), pp. 111-129.

Cornish F. (1990). Anaphore pragmatique, référence, et modèles du discours. *Recherches linguistiques 14*, pp. 81-96.

Cornish F. (1996). Coherence: the lifeblood of anaphora. *Belgian Journal of Linguistics 10*, pp. 37-54.

Cornish F. (2000). L'accessibilité cognitive des référents, le Centrage d'attention et la structuration du discours : une vue d'ensemble. *Verbum 22*(1), pp. 7-30.

Cornish F. (2001). Anaphora, Text, and the Construction of Discourse: A Practical Application. In L. Degand, Y. Bestgen, W. Spooren, L. van Waes (Éds), *Multidisciplinary Approaches to Discourse*, pp. 111-122. Amsterdam / Münster : Stichting Neerlandistiek & Nodus Publikationen.

Cornish F. (2003). The roles of (written) text and anaphor-type distribution in the construction of discourse. *Text 23*(1), pp. 1-26.

Cornish F. (2010). Anaphora: Text-based or discourse-dependent?: Functionalist vs. formalist accounts. *Functions of Language 17*(2), pp. 207-241.

Cruse D. A. (1986). *Lexical Semantics*. Cambridge : Cambridge University Press.

Cusin-Berche F. (1999). Le lexique en mouvement : création lexicale et production sémantique. *Langages 136*, pp. 5-26.

de Mulder W. (1994). Déterminants, cohérence et raisonnement par défaut. *Travaux de linguistique 29*, pp. 93-105.

de Mulder W. (2000). Démonstratifs et accessibilité. *Verbum 22*(1), pp. 103-125.

Delcambre I. & Lahanier-Reuter D. (Éds). (2012). Littéracies universitaires : nouvelles perspectives. *Pratiques 153-154*, http://pratiques.revues.org/1905. Consulté le 15/09/2017.

Denis P. & Sagot B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language resources and evaluation 46*(4), pp. 721-736.

Drouin P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée XII*(2), pp. 45-64.

Dufour F. & Rosier L. (2012). Introduction. Héritages et reconfigurations conceptuelles de l'analyse du discours « à la française » : perte ou profit? *Langage et société 140*(2), pp. 5-13.

Elalouf M.-L. (2011). Constitution de corpus scolaires et universitaires : vers un changement d'échelle ? *Pratiques 149-150*, pp. 56-70.

Falaise A., Tutin A. & Kraif O. (2011). Une interface pour l'exploitation de corpus arborés par des non informaticiens : la plate-forme ScienQuest du projet Scientext. *TAL 52*(3), pp. 241-246.

Fillmore C. J. (1992). « Corpus linguistics » or « Computer-aided arm-chair linguistics ». In J. Svartik (Éd.), *Directions in Corpus Linguistics*, pp. 35-60. Berlin / New-York : Mouton de Gruyter.

Flaux N. & Van de Velde D. (2000). *Les noms en français : esquisse de classement*. Gap : Ophrys.

Florez M. (2014). La citation positionnée dans l'écrit scientifique. In A. Tutin, F. Grossmann (Éds), *L'écrit scientifique : du lexique au discours. Autour de Scientext*, pp. 67–84. Rennes : Presses Universitaires de Rennes.

Frérot C. (2005). *Construction et évaluation en corpus variés de lexiques syntaxiques pour la résolution des ambigüités de rattachement prépositionnel* (Doctorat Nouveau Régime). Université Toulouse II Le Mirail, Equipe de Recherche en Syntaxe et Sémantique, Toulouse.

Frérot C., Bourigault D. & Fabre C. (2003). Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. *TAL 44*(3), pp. 167-186.

Fuchs C. (1986). Le vague et l'ambigu : deux frères ennemis. *Quaderni di Semantica 7*(2), pp. 235-245.

Fuchs C. (1991). Polysémie, interprétation et typicalité : l'exemple de « pouvoir ». In D. Dubois (Éd.), *Sémantique et cognition, catégories, prototypes, typicalité*, pp. 161-170. Paris : Editions du Centre National de la Recherche Scientifique.

Fuchs C. (1997). L'interprétation des polysèmes grammaticaux en contexte. In G. Kleiber, M. Riegel (Éds), *Les formes du sens*, pp. 127-133. Louvain-la-Neuve : Duculot.

Gaatone D. (1999). Réflexions sur la syntaxe de 'ne ... que'. In M. Plénat, M. Aurnague, A. Condamines, J.-P. Maurel, C. Molinier, C. Muller (Éds), *L'emprise du sens Structures linguistiques et interprétations*, pp. 101-115. Amsterdam / Atlanta : Rodopi.

Gasiglia N. (2004). Faire coopérer deux concordanciers-analyseurs pour optimiser les extractions en corpus. *Revue française de linguistique appliquée IX*(1), pp. 45-62.

Goigoux R. (Éd.). (2016). *Etude de l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des premiers apprentissages. Rapport de la recherche « Apprendre à lire et à écrire au Cours Préparatoire »*. Lyon : Institut Français de l'Education.

Goody J. (1979). *La raison graphique*. Paris : Editions de Minuit.

Goody J. (1994). *Entre l'oralité et l'écriture*. Paris : P.U.F.

Grabar N. & Hamon T. (2004). Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'intelligence artificielle 18*(1), pp. 57-85.

Granger S. (2008). Learner Corpora. In A. Lüdeling, M. Kytö (Éds), *Corpus linguistics: an international handbook* (Vol. 1), pp. 259-275. Berlin / New York : Walter de Gruyter.

Granger S. & Paquot M. (2009). Lexical Verbs in Academic Discourse: A Corpus-driven Study of Learner Use. In M. Charles, S. Hunston, D. Pecorari (Éds), *Academic Writing: At the Interface of Corpus and Discourse*, pp. 193-214. London & New York : Continuum.

Gréa P. (2015). Entre et parmi : deux perspectives sur la pluralité. *Travaux de linguistique 70*(1), pp. 7-38.

Gries S. (2012). Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: Towards more and more fruitful exchanges. In M. Huber, J. Mukherjee (Éds), *Corpus Linguistics and Variation in English: Theory and Description*, pp. 41-63. Amsterdam : Rodopi.

Gross G. (1994). Classes d'objets et description des verbes. *Langages 115*, pp. 15-30.

Gross M. (1975). *Méthodes en syntaxe: régime des constructions complétives*. Paris : Hermann.

Grossmann F. (2014). De quelques routines phraséologiques liées aux verbes parenthétiques dans les genres scientifiques (Vol. 8), pp. 759-770. *4e Congrès Mondial de Linguistique Française (CMLF)*, Berlin : SHS Web of Conferences.

Grossmann F. & Tutin A. (2010). Les marqueurs verbaux de constat : un lieu de dialogisme dans l'écrit scientifique. In J. Brès, A. Nowakowska, J.-M. Sarale, S. Sarrazin (Éds), *Actes du colloque international Dialogisme : langue, discours*. Montpellier : Praxiling.

Guilbert L. (1975). *La créativité lexicale*. Paris : Larousse.

Gundel J. K. (1998). Centering Theory and the Givenness Hierarchy: Towards a Synthesis. In M. A. Walker, A. K. Joshi, E. F. Prince (Éds), *Centering Theory in Discourse*, pp. 183-198. Oxford : Clarendon Press.

Gundel J. K., Hedberg N. & Zacharski R. (2000). Statut cognitif et forme des anaphoriques indirects. *Verbum 22*(1), pp. 79-102.

Habert B. (2004). Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs. *Revue française de linguistique appliquée IX*(1), pp. 5-24.

Habert B. (2005). Portrait de linguiste(s) à l'instrument. *Texto! [en ligne] X*(4), http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html. Consulté le .

Habert B. (2009). *Construire des bases de données pour le français: Notions* (Vol. 1). Paris : Ophrys.

Habert B., Nazarenko A. & Salem A. (1997). *Les linguistiques de corpus*. Paris : Armand Colin.

Halliday M. A. K. (1992). Language as system and language as instance: The corpus as a theoretical construct. In J. Svartik (Éd.), *Directions in Corpus Linguistics*, pp. 61-77. Berlin / New-York : Mouton de Gruyter.

Halliday M. A. K. & Hasan R. (1976). *Cohesion in English*. (S.l.) : Longman.

Harris Z. S. (1969). Analyse du discours. *Langages 4*(13), pp. 8-45. Traduction par F. Dubois-Charlier.

Hartwell L. M. & Jacques M.-P. (2012). A Corpus-Informed Text Reconstruction Resource for Learning About the Language of Scientific Abstracts. In L. Bradley, S. Thouësny (Éds), , pp. 117-123. *CALL: Using, Learning, Knowing, EUROCALL Conference, Gothenburg, Sweden, 22-25 August 2012, Proceedings*, Research-publishing.net.

Hartwell L. M. & Jacques M.-P. (2014). Authorial presence in French and English: 'Pronoun + Verb' patterns in Biology and Medicine research articles. *Discours 15*, http://discours.revues.org/8941. Consulté le 01/10/2017.

Hatier S. (à paraitre). Identification et analyse linguistique des noms du LST. In M.-P. Jacques, A. Tutin (Éds), *D'une discipline à l'autre : lexique transversal et formules discursives des sciences humaines*. Londres : ISTE.

Hatier S. (2013). Extraction des mots simples du lexique scientifique transdisciplinaire dans les écrits de sciences humaines : une première expérimentation. In *Actes de la 15e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2013)*, pp. 138–149. Les Sables d'Olonne, France.

Hatier S. (2016). *Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS* (Doctorat Nouveau Régime). Université Grenoble Alpes, Grenoble.

Hatier S., Augustyn M., Yan R., Tran T. T. H., Tutin A. & Jacques M.-P. (2016). French cross-disciplinary scientific lexicon: extraction and linguistic analysis. In T. Margalitadze, G. Meladze (Éds), *Proceedings of the XVII EURALEX International congress Lexicography & Linguistic diversity*, pp. 355–366. Tbilisi : Ivane Javakhishvili Tbilisi State University.

Hearst M. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING-92*, pp. 539-545. Nantes.

Heiden S., Magué J.-P. & Pincemin B. (2010). Une plateforme logicielle open-source pour la textométrie – conception et développement. In S. Bolasco (Éd.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, pp. 1021-1032. Edizioni Universitarie di Lettere Economia Diritto.

Heurley L. (2001). Compréhension et utilisation de textes procéduraux : l'effet de l'ordre de mention des informations. *Revue française de linguistique appliquée 6*(2), pp. 29-46.

Heurley L. & Ganier F. (2006). L'utilisation des textes procéduraux : lecture, compréhension et exécution d'instructions écrites. *Intellectica 44*, pp. 45-62.

Ho-Dac M. (2007). *La position initiale dans l'organisation du discours : une exploration en corpus*. Doctorat Nouveau Régime. Université Toulouse II Le Mirail, Equipe de Recherche en Syntaxe et Sémantique, Toulouse.

Ho-Dac M., Jacques M.-P. & Rebeyrolle J. (2004). Sur la fonction discursive des titres. In S. Porhiel, D. Klingler (Éds), *L'unité texte*, pp. 125-152. Pleyben : Perspectives.

Ho-Dac M. & Péry-Woodley M.-P. (2014). Annotation des structures discursives : l'expérience ANNODIS. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer, S. Prévost (Éds), *4e Congrès Mondial de Linguistique Française* (Vol. 8), pp. 2647-2661. Berlin : SHS Web of Conferences.

Hoek L. H. (1981). *La marque du titre*. La Haye : Mouton.

Illouz G. (2000). *Typage de données textuelles et adaptation des traitements linguistiques application a l'annotation morpho-syntaxique* (Doctorat Nouveau Régime). Université Paris 11.

Jackiewicz A. (1996). L'expression lexicale de la relation d'ingrédience (partie-tout). *Faits de langues 7*, pp. 53-62.

Jackiewicz A. (2004). Les séries linéaires dans le discours : marques, opérations et structures sous-jacentes. In *Journée de l'ATALA : Modéliser et décrire l'organisation discursive à l'heure du document numérique*. La Rochelle.

Jackiewicz A. & Minel J.-L. (2003). L'identification des structures discursives engendrées par les cadres organisationnels. In *TALN'2003* (Vol. 1), pp. 155-164. Batz-sur-Mer.

Jacques M.-P. (1997). Un atelier d'alphabétisation et de lutte contre l'illettrisme en Haute-Garonne. In C. El Hayek (Éd.), *Territoires à livre ouvert La lutte contre l'illettrisme en milieu rural*, pp. 179-182. Paris : La Documentation française.

Jacques M.-P. (2000). La réduction du syntagme terminologique au fil du discours. *Cahiers de Grammaire 25*, pp. 93-114.

Jacques M.-P. (2001). Analyse de corpus pour l'étude de la variation terminologique. *Journée Linguistique de Corpus*, Lorient.

Jacques M.-P. (2002a). Comparaison de corpus pour l'étude de la réduction de termes complexes. *Séminaire ERSS*, Toulouse.

Jacques M.-P. (2002b). Mutations du terme complexe en discours spécialisé. *BULAG 27*, pp. 105-118.

Jacques M.-P. (2003a). Réduction et ambiguïté en discours spécialisé. In E. Hajicová, A. Kotešovcová, J. Mírovský (Éds), *Proceedings of CIL 17, CD-ROM*. Prague : Matfyzpress, MFF UK.

Jacques M.-P. (2003b). Repérage de termes réduits : intérêt et limites de l'analyse distributionnelle, pp. 135-144. *Cinquièmes rencontres Terminologie et Intelligence Artificielle TIA-2003*, Strasbourg : LIIA-ENSAIS.

Jacques M.-P. (2005a). Pourquoi une linguistique de corpus ? In G. Williams (Éd.), *La linguistique de corpus*, pp. 21-30. Rennes : Presses Universitaires de Rennes.

Jacques M.-P. (2005b). Que, la valse des étiquettes. In *Actes de TALN* (Vol. 1), pp. 133-142. Dourdan.

Jacques M.-P. (2005c). Structure matérielle et contenu sémantique du texte écrit. *CORELA (Cognition, Représentation, Langages) 3*(2), https://doi.org/10.4000/corela.560.

Jacques M.-P. (2006). L'emploi de termes réduits comme révélateur de la centralité dans le domaine. In *MOTS, TERMES ET CONTEXTES*, pp. 299-308. Bruxelles.

Jacques M.-P. (2011). Nous appelons X cet Y : X est-il un terme émergent ? In K. Kageura, P. Zweigenbaum (Éds), *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, pp. 31-37. Paris : INALCO.

Jacques M.-P. (2012). Quelle méthodologie pour quel objet ? *Sur le statut et l'utilisation des corpus en linguistique*, Montpellier, France.

Jacques M.-P. (2014). Structure textuelle de l'article scientifique. Les intertitres et la construction rhétorique en sciences humaines et sociales. In A. Tutin, F. Grossmann (Éds), *L'écrit scientifique : du lexique au discours. Autour de Scientext*, pp. 199-215. Rennes : Presses Universitaires de Rennes.

Jacques M.-P. (2016a). D'un corpus à l'identification automatique d'erreurs d'apprenants, pp. 22-29. *Atelier Enseignement des langues et TAL*, Paris.

Jacques M.-P. (2016b). Une linguistique outillée, pour quels objets ? *Histoire Epistémologie Langage 38*(2), pp. 87-99.

Jacques M.-P. (2017a). Intertitres et construction discursive en texte scientifique. In M. Bilger, L. Buscail, F. Mignon (Éds), *Langue française mise en relief. Aspects grammaticaux et discursifs*, pp. 145-158. Perpignan : Presses Universitaires de Perpignan.

Jacques M.-P. (2017b). La structuration textuelle en discours scientifique : comparaison oral / écrit. *CHIMERA: Romance Corpora and Linguistic Studies 4*(1), pp. 89-115.

Jacques M.-P. & Aussenac-Gilles N. (2006). Variabilité des performances des outils de TAL et genre textuel. *T.A.L. 47*(1), pp. 11-32.

Jacques M.-P., Hartwell L. & Falaise A. (2013). Techniques de TAL et corpus pour faciliter les formulations en anglais scientifique écrit, pp. 146-159. *20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*.

Jacques M.-P., Mojahid M. & Sarda L. (2001). Repérer les structures du texte Eléments pour la construction d'un modèle d'analyse, pp. 99-113. *Colloque International sur le Document Electronique*, Toulouse : Europia.

Jacques M.-P. & Poibeau T. (2010). Étudier des structures de discours : préoccupations pratiques et méthodologiques. *CORELA (Cognition, Représentation, Langages) 8*(1), https://doi.org/10.4000/corela.1855.

Jacques M.-P. & Rebeyrolle J. (2005). Quand la linguistique de corpus s'intéresse à une problématique de discours... *4es Journées de la Linguistique de Corpus*, Lorient.

Jacques M.-P. & Rebeyrolle J. (2006). Titres et structuration des documents, pp. 1-12. *International Symposium: Discourse and Document*, Caen : Presses Universitaires de Caen.

Jacques M.-P., Rebeyrolle J. & Ho-Dac M. (2004). Quelques aspects méthodologiques d'une étude de la fonction discursive des titres en corpus. *Journée de l'ATALA : Modéliser et décrire l'organisation discursive à l'heure du document numérique*.

Jacques M.-P. & Rinck F. (2017). Un corpus de "littéracie avancée" : résultat et point de départ. *Corpus 16*, pp. 217-237.

Jacques M.-P. & Soubeille A.-M. (2000). Partages des termes, partage des connaissances ? Construire une modélisation unique de plusieurs corpus, pp. 313-324. *IC'2000 Journées Francophones d'Ingénierie de la connaissance*, Toulouse : Institut de Recherche en Informatique de Toulouse.

Jacques M.-P. & Tutin A. (2015). Termes linguistiques et lexique transdisciplinaire des sciences humaines : observations en contexte. *TOTH (Terminologie et Ontologies, Théories et applications)*, Chambéry, France.

Jacquey E., Tutin A., Kister L., Jacques M.-P., Hatier S. & Ollinger S. (2013). Filtrage terminologique par le lexique transdisciplinaire scientifique : une expérimentation en sciences humaines. In *Proceedings 10th International Conference on Terminology and Artificial Intelligence*, pp. 121-128.

Kara M. (2004). Pratiques de la citation dans les mémoires de maîtrise. *Pratiques 121-122*, pp. 111-142.

Kintsch W. & Van Dijk T. A. (1975). Comment on se rappelle et on résume des histoires. *Langages* (40), pp. 98-116.

Kleiber G. (1986). Adjectif démonstratif et article défini en anaphore fidèle. In J. David, G. Kleiber (Éds), *Déterminants : syntaxe et sémantique*, pp. 169-185. Paris : Klincksieck.

Kleiber G. (1989). Référence indirecte ou de la divergence sur les anaphores divergentes. *Cahiers de Praxématique 12*, pp. 51-74.

Kleiber G. (1990a). Marqueurs référentiels et processus interprétatifs : pour une approche « plus sémantique ». *Cahiers de linguistique française 11*, pp. 241-258.

Kleiber G. (1990b). Quand il n'a pas d'antécédent. *Langages 97*, pp. 24-50.

Kleiber G. (1994). Contexte, interprétation et mémoire : approche standard vs approche cognitive. *Langue Française 103*, pp. 9-22.

Kleiber G. (1997a). Cognition, sémantique et facettes : une « histoire » de livres et de... romans. In G. Kleiber, M. Riegel (Éds), *Les formes du sens*, pp. 219-230. Louvain-la-Neuve : Duculot.

Kleiber G. (1997b). Quand le contexte va, tout va et ... inversement. In C. Guimier (Éd.), *Co-texte et calcul du sens Actes de la table ronde tenue à Caen*, pp. 11-29. Caen : Presses Universitaires de Caen.

Kleiber G. (1997c). Sens, référence et existence : que faire de l'extra-linguistique ? *Langages 127*, pp. 9-37.

Kleiber G. (1999). Il y a contexte et contexte. In M. Plénat, M. Aurnague, A. Condamines, J.-P. Maurel, C. Molinier, C. Muller (Éds), *L'emprise du sens Structures linguistiques et interprétations* (Vol. Collection « Faux titre », 174), pp. 167-181. Amsterdam / Atlanta : Rodopi.

Kleiber G. (2001). Regards sur l'anaphore et la Giveness Hierarchy. In H. Kronning, C. Noréen, B. Novén, G. Ransbo, L.-G. Sundell, B. Svane (Éds), *Langage et référence Mélanges offerts à Kerstin Jonasson à l'occasion de ses soixante ans* (Vol. 63), pp. 310-322. Uppsala : Acta Universitatis Upsaliensis.

Labov W. (1975). *What is a Linguistic Fact?* Lisse : Peter de Ridders Press.

Labov W. (1993). *Le parler ordinaire: la langue dans les ghettos noirs des Ètats-Unis*. Paris : Éd. de Minuit.

Labov W. (1996). When Intuitions Fail. In L. McNair, K. Singer, L. Dolbrin, M. Aucon (Éds), *Papers from the Parasession on Theory and Data in Linguistics* (Vol. 32), pp. 77-106.

Laflamme S. (2007). Analyses qualitatives et quantitatives : deux visions, une même science. *Nouvelles perspectives en sciences sociales 3*(1), p. 141.

Lahire B. (2004). Introduction. In *À quoi sert la sociologie ?*, pp. 5-12. Paris : La Découverte.

Lambrecht K. (1994). *Information structure and sentence form*. Cambridge : Cambridge University Press.

Landragin F. (2011). De la saillance visuelle à la saillance linguistique. In O. Inkova (Éd.), *Saillance : aspects linguistiques et communicatifs de la mise en évidence dans un texte.*, pp. 67-83. Besançon : Presses universitaires de Franche-Comté.

Landragin F. (2012). La saillance : questions méthodologiques autour d'une notion multifactorielle. *Faits de langues 39*, pp. 15-31.

Landragin F. (2016). Conception d'un outil de visualisation et d'exploration de chaînes de coréférences. *Journées d'Analyse de Données Textuelles*, Nice.

Le Pesant D. & Mathieu-Colas M. (1998). Introduction aux classes d'objets. *Langages 131*, pp. 6-33.

Lecolle M. (2003). *Métonymies et figures de référenciation dans la presse écrite généraliste analyse sémantique et rhétorique* (Doctorat Nouveau Régime). Université Toulouse II Le Mirail, Equipe de Recherche en Syntaxe et Sémantique, Toulouse.

Lecourt D. (2015). *La philosophie des sciences* (Vol. 6e éd.). Paris : Presses Universitaires de France.

Leech G. (1992). Corpora and theories of linguistic performance. In J. Svartik (Éd.), *Directions in Corpus Linguistics*, pp. 105-122. Berlin / New-York : Mouton de Gruyter.

Léglise I., Canut E., Desmet I. & Garric N. (2006). Applications et implications en sciences du langage : Introduction. In I. Léglise, E. Canut, I. Desmet, N. Garric (Éds), *Applications et implications en sciences du langage*, pp. 9-15. Paris : L'Harmattan.

Lemarié J., Lorch R. F., Eyrolle H. & Virbel J. (2008). SARA: A Text-Based and Reader-Based Theory of Signaling. *Educational Psychologist 43*(1), pp. 27-48.

Lemarié J., Lorch R. F. & Péry-Woodley M.-P. (2012). Understanding how headings influence text processing. *Discours 10*, https://doi.org/10.4000/discours.8600.

Léon J. (2005). Claimed and Unclaimed Sources of Corpus Linguistics. *Henry Sweet Society Bulletin 44*, pp. 36-50.

Léon J. (2008). Aux sources de la « Corpus Linguistics » : Firth et la London School. *Langages 171*, pp. 12-33.

Léon J. (2015). Linguistique appliquée et traitement automatique des langues. Étude historique et comparative. *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle 12*(3), https://doi.org/10.4000/rdlc.949.

Leroy S. (2001). *Entre identification et catégorisation, l'antonomase du nom propre en français* (Doctorat Nouveau Régime). Université Paul Valéry, Montpellier.

Leroy S. (2004). Extraire sur patrons : allers et retours entre analyse linguistique et repérage automatique. *Revue française de linguistique appliquée 9*(1), pp. 25-43.

L'Homme M.-C. (2005). Sur la notion de « terme ». *Meta: Journal des traducteurs 50*(4), pp. 1112-1132.

Libersan L., Claing R. & Foucambert D. (2010). *Stratégies d'écriture dans les cours de la formation spécifique. Rapport 2009-2010.* Québec, Canada : CCMD.

Lopez C. (2013). *Titrage automatique de documents textuels*. Sarrebruck : Éditions universitaires européennes.

Lorch J. Robert F. & Lorch E. P. (1996). Effects of Headings on Text Recall and Summarization. *Contemporary Educational Psychology 21*(3), pp. 261-278.

Luc C. & Virbel J. (2001). Le modèle d'architecture textuelle Fondements et expérimentation. *Verbum 23(1)*, pp. 103-123.

Lüdeling A. & Kytö M. (Éds). (2008). *Corpus linguistics: an international handbook* (Vol. 1). Berlin / New York : Walter de Gruyter.

Maingueneau D. (2005). L'analyse du discours et ses frontières. *Marges Linguistiques 9*, pp. 64-75.

Malrieu D. (2004). Linguistique de corpus, genres textuels, temps et personnes. *Langages 153*, pp. 73-85.

Malrieu D. & Rastier F. (2001). Genres et variations morphosyntaxiques. *Traitement Automatique des Langues 42*(2), pp. 547-577.

Mann W. C. & Thompson S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text 8*(3), pp. 243-281.

Masseron C. (2004). Les paradoxes de CAR, entre énoncés et discours – ou les difficultés d'un traitement didactique des connecteurs. *Linx* (51), pp. 107-127.

Matthiessen C. & Halliday M. A. K. (2009). *Systemic Functional Grammar: A First Step into the Theory*. China : Higher Education Press.

McEnery T., Xiao R. & Tono Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London / New York : Routledge.

Mel'cuk I. (1997). *Vers une linguistique Sens-Texte*. Collège de France.

Otman G. (1995). Introduction. *La Banque des mots (Paris)* (NS7), pp. 3-4.

Paroubek P., Pouillot L.-G., Robba I. & Vilnat A. (2005). EASy : campagne d'évaluation des analyseurs syntaxiques (Vol. 2), pp. 3-12. *TALN'2005*, Dourdan.

Pecman M. (2007). Approche onomasiologique de la langue scientifique générale. *Revue française de linguistique appliquée XII*(2), pp. 79-96.

Péry-Woodley M.-P. (1995). Quels corpus pour quels traitements automatiques ? *Traitement Automatique des Langues 36*(1-2), pp. 213-232.

Péry-Woodley M.-P. (2001). Modes d'organisation et de signalisation dans des textes procéduraux. *Langages 141*, pp. 28-46.

Petitjean A. (1989). Les typologies textuelles. *Pratiques* (62), pp. 86-125.

Pilorgé J.-L. (2010). Un lieu de tension entre posture de lecteur et posture de correcteur : les traces des enseignants de français sur les copies des élèves. *Pratiques 145-146*, pp. 85-103.

Pincemin B. (2007). Concordances et concordanciers : de l'art du bon KWAC. In F. Rastier, M. Ballabriga, C. Duteil-Mougel, B. Fouqié (Éds), *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation : actes du XXVIIe colloque d'Albi Langages et signification*, pp. 33-42. Albi, France : CALS-CPST.

Plénat M., Lignon S., Serna N. & Tanguy L. (2002). La conjecture de Pichon. *Corpus* (1), pp. 105-150.

Pontille D. (2007). Matérialité des écrits scientifiques et travail de frontières : le cas du format IMRAD. In P. Hert, M. Paul-Cavallier (Éds), *Sciences et frontières*, pp. 229-253. Fernelmont : E.M.E.

Rastier F. (1994). *Sémantique pour l'analyse*. Paris : Masson - Sciences Cognitives.

Rastier F. (1995). Le terme : entre ontologie et linguistique. *La Banque des mots numéro spécial 7*, pp. 35-65.

Rastier F. (2005). Discours et texte. *Texto!* http://www.revue-texto.net/Reperes/Themes/Rastier_Discours.html. Consulté le 01/09/2017.

Rebeyrolle J. (2000). *Forme et fonction de la définition en discours* (Doctorat Nouveau Régime). Université Toulouse II Le Mirail, Equipe de Recherche en Syntaxe et Sémantique, Toulouse.

Rebeyrolle J., Jacques M.-P. & Péry-Woodley M.-P. (2009). Titres et intertitres dans l'organisation du discours. *Journal of French Language Studies 19(2)*, pp. 269-290.

Rebeyrolle J. & Tanguy L. (2001). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire 25*, pp. 153-174.

Récanati F. (2001). Déstabiliser le sens. *Revue internationale de philosophie 216*, pp. 197-208.

Reichler-Béguelin M.-J. (1988). Anaphore, cataphore et mémoire discursive. *Pratiques 57*, pp. 15-43.

Reichler-Béguelin M.-J. (1995). Déterminant zéro et anaphore. *TRANEL 23*, pp. 177-201.

Riegel M., Pellat J.-C. & Rioul R. (2009). *Grammaire méthodique du français* (4ème). Paris : P.U.F. - Quadrige.

Rondelli F. (2010). Comment les enseignants construisent-ils un objet de savoir ? Exemple de la cohérence textuelle. *Repères 42*, pp. 63-81.

Rowley-Jolivet E. & Carter-Thomas S. (2005). Scientific conference Englishes: Epistemic and Language Community Variations. In G. Cortese, A. Duszak (Éds), *Identity, community, discourse: English in intercultural settings*, pp. 295-320. Bern ; New York : Peter Lang.

Sablayrolles J.-F. (2000). *La néologie en français contemporain*. Paris : Honoré Champion.

Schmid H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *NEMLAP*.

Schnedecker C. & Landragin F. (2014). Les chaînes de référence : présentation. *Langages 195*(3), pp. 3-22.

Séguéla P. (1999). Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés. *Terminologies nouvelles 19*, pp. 52-60.

Siepmann D. (2007). Les marqueurs de discours polylexicaux en français scientifique. *Revue française de linguistique appliquée XII*(2), pp. 123-136.

Slodzian M. (1993). La V.G.T.T. (Vienna General Theory of Terminology) et la Conception Scientifique du monde. *Le langage et l'homme 28(4)*(Socioterminologie), pp. 223-232.

Slodzian M. (1995). Comment revisiter la doctrine terminologique aujourd'hui ? *La Banque des mots numéro spécial 7*, pp. 11-18.

Sollaci L. B. & Pereira M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association 92*(3), pp. 364-371.

Stubbs M. (2001). *Words and phrases: corpus studies of lexical semantics*. Oxford [England] ; Malden, MA : Blackwell Publishers.

Sullet-Nylander F. (2002). Titres de presse et polyphonie. *Romansk Forum 16*(2), pp. 767-775.

Sumpf J. & Dubois J. (1969). Problèmes de l'analyse du discours. *Langages 13*, pp. 3-7.

Swales J. (1990). *Genre analysis: English in academic and research settings*. Cambridge / New York : Cambridge University Press.

Tanguy L. & Hathout N. (2007). *Perl pour les linguistes: programmes en Perl pour exploiter les données langagières*. Paris : Hermès science publications.

Taylor C. (2008). What is corpus linguistics? What the data says. *International Computer Archive of Modern and Medieval English Journal 32*, pp. 179-200.

Teubert W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics 10*(1), pp. 1-13.

Tognini-Bonelli E. (2001). *Corpus Linguistics at Work*. Amsterdam : John Benjamins.

Tran T. T. H. (2014). *Description de la phraséologie transdisciplinaire des écrits scientifiques et réflexions didactiques pour l'enseignement à des étudiants non-natifs : application aux marqueurs discursifs*. Doctorat Nouveau Régime. Université Grenoble 3, Grenoble.

Tutin A. (2007). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée XII*(2), pp. 5-14.

Urieli A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Doctorat Nouveau Régime. Université Toulouse II Le Mirail, Equipe de Recherche en Syntaxe et Sémantique.

Urieli A. (2015). Stratégies pour l'étiquetage et l'analyse syntaxique statistique de phénomènes difficiles en français : études de cas avec Talismane. *TAL 56*(1), pp. 65-89.

Van Dijk T. A. (1995). A Brief Personal History of the Kintsch-van Dijk Theory. In C. Weaver, S. Mannes, C. R. Fletcher (Éds), *Discourse comprehension: Essays in honor of Walter Kintsch*, pp. 383-410. Hillsdale, NJ : Erlbaum.

Vergne J. & Giguet E. (1998). Regards Théoriques sur le « Tagging ». In *TALN'98*, pp. 22-31. Paris.

Véronis J. (2001). Sense tagging: does it make sense ? In *Corpus Linguistics*. Lancaster.

Victorri B. (1997). La polysémie : un artefact de la linguistique ? *Revue de Sémantique et Pragmatique 2*, pp. 41-62.

Victorri B. & Fuchs C. (1996). *La polysémie Construction dynamique du sens*. Paris : Hermès.

Virbel J., Schmid S., Carrio L., Dominguez C., Péry-Woodley M.-P., Jacquemin C., … Garcia-Debanc C. (2005). Approches cognitives de la spatialisation du langage : le cas de l'énumération. In C. Thinus-Blanc, J. Bullier (Éds), *Agir dans l'espace*, pp. 233-254. Paris : Éd. de la Maison des sciences de l'Homme.

Walker M. A. (1998). Centering, Anaphora Resolution, and Discourse Structure. In M. A. Walker, A. K. Joshi, E. F. Prince (Éds), *Centering Theory in Discourse*, pp. 401-435. Oxford : Clarendon Press.

Walker M. A. (2000). Vers un modèle de l'interaction du Centrage avec la structure globale du discours. *Verbum 22*(1), pp. 31-58.

Werth P. (1999). *Text worlds: Representing conceptual space in discourse*. Londres : Longman.

Williams G. (2006). La linguistique et le corpus : une affaire prépositionnelle. *Texto! [en ligne]* pp. 151-158.

Winston M., Chaffin R. & Herrmann D. (1987). A Taxonomy of Part-Whole Relations. *Cognitive Science 11*, pp. 417-444.