Affinités syntaxiques et sémantiques entre mots : apports mutuels de la linguistique et du TAL

Cécile Fabre

Université de Toulouse

Mémoire présenté pour l'obtention d'une habilitation à diriger des recherches

Spécialité : linguistique

Affinités syntaxiques et sémantiques entre les mots – apports mutuels de la linguistique et du traitement automatique des langues

Cécile FABRE

laboratoire clle-erss (umr 5263)

UNIVERSITÉ TOULOUSE 2 - LE MIRAIL ET CNRS

2010

Remerciements

Merci à Marie-Paule Péry-Woodley, qui a dirigé ce travail. J'ai eu beaucoup de chance d'exercer mon métier en collaboration étroite avec elle dont le souci de l'engagement collectif s'accompagne de propositions innovantes et exigeantes sur le plan de l'enseignement et de la recherche.

Je remercie Benoît Habert d'avoir accepté de renouer avec cet exercice académique à l'occasion de ma soutenance. Les échanges réguliers que nous avons eus pendant les mois qui ont précédé m'aident à effectuer le difficile et salutaire « pas de côté ». Merci pour le rôle déterminant et la présence stimulante depuis toujours.

Merci à Adeline Nazarenko et à Alain Polguère d'avoir accepté d'accompagner ce travail en tant que rapporteurs. J'ai beaucoup apprécié leur façon de questionner en profondeur les options que j'ai prises. Leurs suggestions m'aident à prendre du champ et ouvrent d'autres perspectives pour la suite.

Merci à Paola Merlo et Agnès Tutin. Leur participation au jury apporte un éclairage précieux en traitement automatique des langues et linguistique de corpus. Merci à Nabil Hathout d'avoir également accepté le rôle d'examinateur et de m'avoir aidée à formuler plus nettement certaines orientations de ma recherche. Merci pour la constante bienveillance avec laquelle il a répondu à mes nombreuses demandes d'aide dans la réalisation matérielle de ce document.

***

J'ai la chance d'avoir été soutenue et encouragée dans la réalisation de ce travail par de nombreux collègues à l'Université de Toulouse 2 - Le Mirail.

Je remercie Didier Bourigault pour ses remarques sur un document qui lui doit beaucoup : plusieurs pistes explorées ici trouvent leur origine dans le cadre des travaux que j'ai réalisés avec lui autour de SYNTEX. Merci pour cette collaboration généreuse au cours de laquelle j'ai beaucoup appris.

Merci à mes très proches collègues, Ludovic Tanguy et Josette Rebeyrolle, pour leurs relectures et leurs encouragements, et bien au-delà, pour le plaisir du travail réalisé ensemble, moteur principal de mon désir d'avancer dans cette carrière d'enseignant-chercheur. Merci

à Anne Le draoulec pour nos discussions qui m'ont aidée à donner plus de sens à toute cette affaire.

Merci à Franck Sajous pour sa compagnie amicale et secourante, à Myriam Bras, Andrée Borillo, Fabio Montermini pour leurs encouragements et leurs conseils.

Merci à Clémentine Adam, Fanny Lalleman et François Morlane-Hondère. C'est une chance d'entamer mon expérience de direction de thèse avec eux.

Je voudrais également associer à ces remerciements les collègues de l'équipe CLLE-ERSS avec qui j'ai collaboré à divers titres, en particulier Anne Condamines, Mai Ho-Dac, Anna Kupść et Aurélie Picton, ainsi que les collègues du département de sciences du langage pour le travail commun sur le terrain de l'enseignement, en particulier Barbara Köpke et Christiane Soum.

Je ne peux ici faire état de l'ensemble des collaborations qui ont nourri les travaux dont je rends compte et que j'ai nouées avec des collègues extérieurs à l'Université de Toulouse 2 - Le Mirail. J'envoie simplement un salut amical à Pascale Sébillot, qui m'a donné la chance de démarrer cette aventure.

***

J'ai écrit mon HdR pendant les périodes de vacances. Cela n'aurait pas été possible sans une solide solidarité familiale. Merci à mes parents et beaux-parents d'avoir joué leur rôle de grands-parents encore plus remarquablement que d'habitude. Merci à Alice et Marianne de n'avoir que modérément protesté contre mon manque de disponibilité. Merci à Philippe, la clef de voûte de l'ensemble.

Table des matières

Présentation : parcours et problématiques de recherche

Je fais ici le bilan des travaux que j'ai menés depuis ma thèse, sur la période allant de 1997 (date de mon recrutement à Toulouse) à 2010, dans le domaine de la linguistique et du Traitement Automatique des Langues (désormais TAL), en profitant de l'occasion de ce rapport de recherches pour faire le point sur la question du lien entre ces deux domaines. Je considérerai ce lien comme mutuel, puisque j'examinerai à la fois la question de l'apport de la linguistique au TAL et du gain que la linguistique peut escompter à mettre à profit les techniques et les méthodes du TAL pour étudier le langage. Dans ce préambule, je commence par retracer rapidement mon parcours, avant d'exposer le cadre et les thématiques de recherche que je vais présenter, ainsi que la façon dont ce document s'organise.

Les différents temps de mon parcours Après trois années de classes préparatoires littéraires, c'est le cours de linguistique générale que propose Robert Martin dans le cadre de la licence de lettres de Paris 4 qui me fait découvrir le champ de la linguistique et m'en donne le goût (en 1990). C'est également Robert Martin qui, signalant l'existence d'une formation en linguistique et informatique, m'ouvre des perspectives de formation toutes nouvelles. Je rejoins l'année suivante la filière dirigée par Laurence Danlos à Paris 7, me convertissant ainsi définitivement à la linguistique et au traitement automatique des langues. Je rencontre Benoît Habert à la fin de la maîtrise et je fais mon mémoire de DEA sous sa direction, sur la question de la représentation du sens des composés nominaux. Un financement de thèse sur cette thématique est proposé à l'Université de Rennes 1, et je prends contact avec Pascale Sébillot en 1993, qui va encadrer mon travail au sein d'une équipe d'intelligence artificielle du laboratoire IRISA jusqu'en 1996. A la fin de ma thèse, je suis recrutée comme ATER à l'ENS de Fontenay-Saint-Cloud où je reprends la collaboration avec B. Habert, auprès de qui j'apprends à travailler sur les corpus. Je rejoins l'année suivante (en 1997) l'Université du Mirail et l'ERSS grâce à un poste de linguistique profilé corpus et TAL, avec pour première mission de mettre en place avec Marie-Paule Péry-Woodley une filière TAL au sein du département de sciences du langage.

J'ai donc eu la chance de suivre un parcours à la fois varié et cohérent dans le domaine du TAL, puisque j'ai pu mener des projets avec une certaine continuité, tout en travaillant dans des contextes différents. A l'IRISA, j'ai découvert l'organisation de la recherche dans un grand laboratoire d'informatique, qui offre un environnement stimulant, même si les thématiques en lien avec la linguistique sont très circonscrites. A l'ERSS, dirigée à l'époque par Marc Plénat, puis par Jacques Durand et aujourd'hui Fabio Montermini, j'ai rejoint une équipe de passionnés des langues et du langage, tournée vers les approches empiriques et quantitatives des phénomènes langagiers, au sein de laquelle, au fil des ans, la thématique TAL va se renforcer. Cette familiarité avec deux versants du monde de l'Université et de la recherche (Université de sciences et laboratoire d'informatique à Rennes, Université de lettres et sciences humaines et laboratoire de linguistique à Toulouse) a certainement façonné ma conception de la recherche, confortée par le champ que j'explore, le TAL, en position d'interface entre ces deux communautés. Au risque de schématiser, je dirais que le versant informatique m'a appris la nécessité et la richesse du travail collaboratif, l'intérêt des enjeux applicatifs. Le versant linguistique - et plus généralement la découverte de la recherche dans une grande université de lettres et sciences humaines - m'a convaincu que les pistes de travail plus personnelles peuvent s'avérer également fertiles. L'encadrement d'étudiants me semble être au cœur de ce compromis.

Le développement du TAL à l'ERSS et au département de sciences du langage Je rejoins tout d'abord l'opération de recherche dirigée par Anne Condamines, qui prend le nom de « Sémantique et corpus » et promeut des approches outillées pour l'étude du sens dans les corpus (Condamines, 2000). Je me trouve immédiatement très à l'aise dans un groupe qui place au cœur de sa réflexion les questions théoriques et méthodologiques qui m'intéressent : la place du corpus et des outils dans l'analyse sémantique, la validation et la généralisation des résultats (*op. cit.*, 24 et sq.) Les approches TAL se sont développées progressivement dans le laboratoire grâce à l'arrivée de plusieurs chercheurs et enseignants-chercheurs - L. Tanguy, D. Bourigault et N. Hathout à la fin des années 1990, plus récemment B. Gaume et A. Kupść, ainsi que F. Sajous, ingénieur. En 2006, nous sommes quelques-uns (D. Bourigault, N. Hathout, M.-P. Péry-Woodley, L. Tanguy et moi-même), à décider de créer un axe TAL spécifique, dirigé par D. Bourigault puis par L. Tanguy. La constitution d'un groupe de recherche mieux identifié nous permet de mettre en avant la spécificité d'une approche des données linguistiques basée sur des traitements à grande échelle à partir de corpus et le recours massif à des outils, ainsi qu'une vision expérimentale du TAL au service de l'analyse linguistique.

Aujourd'hui, j'estime avoir contribué avec mes collègues à la mise en place d'un environnement particulièrement favorable au développement de recherches en linguistique dans le domaine du TAL, qui conjugue, du côté du laboratoire, une équipe TAL dont la constitution comme axe spécifique a favorisé la visibilité et la cohérence, et du côté de l'enseignement, une filière TAL, coordonnée par M.-P. Péry-Woodley (qui la dirige), L. Tanguy et moi-même, qui s'est consolidée au niveau du doctorat et de l'offre en master, même si sa situation reste précaire sur le plan des effectifs étudiants. Nous avons bénéficié en cela d'un contexte local propice : un laboratoire de linguistique dynamique, largement intéressé

par les approches empiriques et quantitatives des données linguistiques ; des collaborations suivies avec les informaticiens de l'IRIT qui travaillent sur le langage (équipes LILaC et IC3).

Linguistique et TAL Je travaille en tant que linguiste dans le domaine du TAL, au niveau de l'enseignement comme de la recherche, et dans une double mesure :

- je m'intéresse au TAL dans ses visées applicatives ;
- j'intègre les outils de TAL dans mon travail d'analyse linguistique, avec pour point de départ les corpus de textes annotés.

Ces deux aspects correspondent à deux façons d'envisager, selon moi, un rapport fructueux entre TAL et linguistique.

Tout d'abord, le TAL – et de façon générale, les applications relevant du domaine de l'ingénierie des langues – requiert des descriptions linguistiques nouvelles, sollicitées par les problématiques et les objets particuliers que l'on veut traiter. La confrontation avec les objectifs et les applications du TAL a souvent montré le caractère inadapté de ressources conçues hors de la référence à ce cadre, et doit de ce fait solliciter des études ancrées dans les problématiques de ce champ. En conséquence, le TAL n'est pas un horizon hypothétique et lointain pour le linguiste qui s'intéresse à ce champ, il doit constituer son terrain d'étude. Ainsi, je m'intéresse au développement d'applications de TAL qui font appel à des méthodes et des connaissances linguistiques, et je fais en sorte de situer certaines de mes contributions dans ce cadre applicatif, de manière à vérifier l'utilisabilité des descriptions produites. J'explore parallèlement l'idée selon laquelle les applications du TAL contribuent à renouveler les objets d'étude de la linguistique et élargissent le champ de ses questionnements.

Par ailleurs, je considère également que la linguistique peut bénéficier pour elle-même des méthodes et des outils du TAL, lequel « vise dans une certaine mesure à doter la linguistique d'une instrumentation » (Valette, 2010). Cette perspective va dans le sens d'une linguistique outillée (Habert, 2005), plus précisément, elle met en avant l'intérêt d'un outillage avancé qui s'appuie sur des procédures automatiques de découverte, basées principalement sur l'annotation automatique de textes assortie de techniques statistiques adaptées aux besoins de la description linguistique.

En ce sens, ce que je veux signifier c'est que faire du TAL est une façon particulière de faire de la linguistique : je m'efforce de travailler à cette intégration, et ce rapport fait état de mes activités et de mes réflexions en ce sens [1]. La question du rôle de la linguistique dans le TAL est centrale aussi bien dans mon travail de recherche que dans mes missions d'enseignante, puisque la filière TAL forme des étudiants de linguistique qui sont amenés à travailler dans ce domaine, avec des compétences qu'ils ne pourraient guère faire valoir

1. Je continue néanmoins à parler séparément de linguistique et de TAL, en termes de coopération entre deux champs de recherche, d'abord parce que je n'ai pas trouvé de manière satisfaisante de désigner cette linguistique du TAL ou ce TAL linguistique, comme on parle de linguistique de corpus. Ensuite, parce qu'à l'heure actuelle, cette intégration est loin d'être acquise et que l'on pense plus couramment en termes de recours par des linguistiques à des méthodes de TAL, ou de recours à des linguistes dans des projets de TAL.

si le rôle des traitements linguistiques s'avérait marginal dans les applications du TAL.

Problématiques de recherche Au sein de ce cadre général, les travaux que je mène ont porté principalement sur deux thématiques de recherche.

Je me suis intéressée, sur l'ensemble de la période considérée, et dans la continuité de ma thèse, à la question du repérage des rapports d'équivalence ou de proximité sémantique entre différents types d'unités - mots, termes, structures prédicatives. Il s'agit bien d'un enjeu pour le TAL : identifier sous la diversité des formulations des éléments de contenu similaire est un objectif crucial pour de nombreuses applications qui visent l'accès à l'information dans les textes. En retour, ce besoin applicatif m'a amenée à retravailler avec d'autres objectifs et d'autres catégories des questions linguistiques fondamentales comme la paraphrase et les relations sémantiques, au niveau lexical d'abord, aux niveaux des phrases et des textes ensuite.

Impliquée aux côtés de D. Bourigault dans le projet de développement et d'exploitation d'un analyseur syntaxique automatique, SYNTEX, je me suis également intéressée à une question qui est au cœur des problèmes d'ambiguïté syntaxique, à savoir le rattachement des groupes prépositionnels. J'ai travaillé en particulier à la mise au point d'une méthode permettant de distinguer des types différents de rattachement prépositionnel, de nature argumentale ou adjonctive. Dans ce cas également, mon travail est guidé par un objectif qui relève du TAL, puisqu'il s'agit à l'origine d'améliorer les performances de l'analyseur. Mais en retour, ce projet m'a amené à retravailler avec d'autres méthodes une question linguistique centrale en syntaxe, qui est la distinction entre arguments et circonstants, et de développer des méthodes qui renouvellent la manière de l'examiner.

Ces deux versants de mon travail sont en apparence disjoints. Plusieurs connexions existent néanmoins entre les deux. La première peut paraître relativement anecdotique, mais pour moi elle a été déterminante : les travaux de recherche que j'ai menés sur la question de la proximité sémantique utilisent des programmes qui traitent des corpus analysés syntaxiquement. Un premier projet réalisé en collaboration avec B. Habert m'avait convaincue de la nécessité de disposer d'un analyseur syntaxique à large couverture, c'est la raison pour laquelle j'ai souhaité accompagner D. Bourigault dans le projet SYNTEX. Ce cheminement atteste de l'importance que revêt selon moi un outillage adéquat dans l'analyse linguistique. L'autre connexion est également de nature méthodologique. A travers ces deux thématiques, qui concernent l'étude des relations syntaxiques et sémantiques, je poursuis une démarche semblable, qui consiste à analyser les effets d'une approche inductive à partir de corpus sur les catégories de description usuelles en linguistique – relations lexicales dans un cas, types de compléments dans l'autre. Dans les deux cas, je cherche à mesurer l'apport pour la linguistique d'un dispositif alliant l'exploration de grands corpus en français et des outils de TAL dédiés à leur annotation et à l'extraction et l'analyse des données. Cette démarche d'analyse linguistique fondée sur les corpus et qui fait appel à des méthodes de TAL a pour effet d'élargir le champ d'investigation et de repenser des catégorisations existantes : travailler sur l'entre-deux plutôt que sur les pôles bien identifiés de l'opposition entre argument et circonstant, travailler sur des relations de sens qui ne relèvent pas des relations lexicales traditionnellement recensées. D'où le terme d'affinités

pour lequel j'ai opté dans le titre de ce rapport.

Organisation du document Je présente successivement mes deux principaux objets de recherche : les travaux autour de la proximité sémantique puis ceux portant sur l'étude des mécanismes de complémentation. Dans ces deux parties, je pose les termes de la question du point de vue linguistique, et je montre pourquoi le TAL sollicite des travaux sur ces questions et en quoi il conduit à les reformuler. En décrivant ma contribution sur ces thématiques j'essaie de montrer en quoi ce travail, qui consiste à doter l'étude linguistique de nouveaux outils d'exploration, déplace les conditions d'observation et dégage de nouveaux objets de description. Je dresse dans une troisième partie un bilan d'orientation méthodologique, dans lequel je cherche à montrer, à la lumière des expériences présentées dans les deux premières parties, à quelles conditions peut s'opérer une coopération fructueuse entre linguistique et TAL, et dans quelle mesure les méthodes d'investigation linguistique évoluent et se diversifient au contact du TAL et, de façon plus générale, par l'intégration d'outils informatiques de calcul et d'analyse.

Voici le détail du déroulement chapitre par chapitre.

Dans le chapitre 1, je montre de quelle manière le TAL peut s'emparer d'un objet linguistique – ici, l'équivalence sémantique – et proposer d'autres manières de le considérer. Dans le chapitre 2, je fais état de deux expériences menées au début de la période que je considère. Elles sont décrites en détail de manière à pouvoir y revenir dans la partie finale, car elles font état d'une manière d'organiser le triptyque linguistique, corpus et TAL qui est au cœur de mes travaux. Dans le chapitre 3, je présente mes travaux et réflexions autour de l'analyse distributionnelle automatique et de l'exploitation en linguistique et en TAL de la ressource lexicale constituée en aval du traitement. Le chapitre 4 clôt la première partie par la présentation d'un élargissement récent du champ d'investigation, du lexique vers le discours.

La deuxième partie, consacrée aux études sur la complémentation, commence par un court chapitre qui montre comment la participation au développement de l'analyseur syntaxique automatique SYNTEX a initié et inspiré les travaux que j'ai menés sur la complémentation verbale. Le chapitre 6 présente la méthode que j'ai proposée pour étudier l'opposition argument / circonstant avec des méthodes quantifiées de type *corpus-driven*. Le chapitre 7 présente les résultats de ce travail du point de vue d'une analyse descriptive de la complémentation verbale.

La dernière partie expose la double préoccupation qui est la mienne et que j'ai indiquée plus haut : défendre la contribution de la linguistique dans le TAL (chapitre 8) et promouvoir une linguistique qui s'empare des instruments du TAL (chapitre 9). Je conclus ce mémoire en évoquant les pistes de travail et les projets d'encadrement que je souhaite développer pour la suite.

Première partie

Détecter des liens de proximité sémantique à partir des corpus

Présentation de la première partie

Savoir repérer et mettre en relation des formulations distinctes d'une même « idée » est un objectif partagé par la plupart des applications du TAL. L'enjeu est de réussir à déterminer qu'on a affaire à un même contenu informationnel, exprimé à travers des réalisations linguistiques diversifiées. Le développement des applications visant un accès précis au contenu des textes et supposant de ce fait la prise en compte de la variété des formulations possibles d'un même contenu d'information, a relancé et renouvelé les travaux autour de la question de la paraphrase et de l'équivalence sémantique, donnant ainsi à cette problématique centrale de la linguistique un ancrage pratique et applicatif. Si le TAL s'est d'abord intéressé à la paraphrase dans le but de pouvoir générer automatiquement des variantes textuelles (Stede 1996), de nombreux travaux concernent depuis quelques années le repérage et l'acquisition de segments textuels liés par des relations de paraphrase ou de similitude sémantique pour mettre en correspondance des passages de textes. La question de la variabilité langagière et du recouvrement sémantique (*semantic overlap*) entre segments textuels sont à l'origine de conférences et de campagnes spécifiques [2] :

> « Many natural language processing applications (...) need a model for this variability phenomenon [= la variabilité de l'expression sémantique] in order to recognize that a particular target meaning can be inferred from different text variants ». (Dagan *et al.*, 2006).

La question de la proximité sémantique y est ainsi posée dans les termes généraux qui sont ceux de la linguistique - paraphrase et inférence - mais en étendant très largement les phénomènes que ces notions recouvrent. On est bien loin du périmètre défini par l'approche linguistique traditionnelle (section 1.1.1). Le champ considéré renvoie à des phénomènes très variés, dans la mesure où on s'intéresse à des zones d'étendue variable (du mot au texte), à des rapports de sens de diverse nature – dont la paraphrase logique n'est qu'un cas très particulier – et à des objectifs applicatifs divers (rapprocher deux textes quelconques, deux traductions, une question et sa réponse potentielle, un texte et son résumé, etc.).

Cette première partie est consacrée à l'exposé des activités que j'ai menées autour de cette question de la proximité sémantique. Dans le premier chapitre, je fais le point sur la question de l'équivalence sémantique en linguistique et en TAL, en précisant comment les objectifs d'une sémantique appliquée amènent à étendre le champ d'investigation concernant les questions de paraphrase et de relations de proximité de sens entre mots. Je m'intéresse en particulier à la façon dont le TAL suscite la prise en compte de relations de nature plus lâche que celles qui sont classiquement identifiées en linguistique, et, dans cette perspective, j'explore un type particulier de relations, à savoir les relations intercatégorielles qui rapprochent les noms et les verbes. Le chapitre suivant présente deux contributions sur ce point : la première expérience vise à rapprocher des termes nominaux et leurs variantes verbales, la deuxième à acquérir des couples nom-verbe qui présentent un lien sémantique fort. Je présente ensuite, dans le chapitre 3, les travaux que j'ai menés dans le cadre de l'approche distributionnelle, qui vise à induire les rapprochements sémantiques

2. En particulier les campagnes RTE (*Recognizing Textual Entailment Challenges*) depuis 2004.

entre mots sur la base de la similarité de leur comportement syntaxique. Je reviens à cette occasion sur la caractérisation des liens nom-verbe mais explore plus généralement la question des relations sémantiques qualifiées de « non classiques ». Dans le dernier chapitre, je présente mes pistes de recherche actuelles, qui visent à étendre, à partir d'indices lexicaux, le repérage de la proximité sémantique au niveau du texte.

Chapitre 1

Rapprochements sémantiques : de l'équivalence à la proximité

Ce chapitre explique comment le fait de considérer la question linguistique fondamentale de la paraphrase depuis le point de vue applicatif du TAL élargit les limites des phénomènes que l'on considère et fournit de nouveaux critères pour leur mise au jour. Il me permet à la fois d'expliquer ma démarche générale – un travail de linguistique alimenté par des questionnements appliqués – et de commencer à présenter les notions sur lesquelles j'ai travaillé sur ce premier volet des affinités sémantiques entre mots – relations non classiques, relations de sens entre mots relevant de catégories grammaticales différentes.

1.1 Conceptions de la paraphrase en linguistique et en TAL

1.1.1 Approches linguistiques de la paraphrase

C. Fuchs définit la paraphrase comme :

> « (...) une relation virtuelle unissant toutes les formulations-types appartenant à un même ensemble paraphrastique $\{F_i\}$, c'est-à-dire – en première approximation – reliées à un même contenu C. » (Fuchs, 1994, p. 44)

Elle est traditionnellement inscrite dans le périmètre de la phrase (ou, si l'on préfère, de la proposition) :

> « presque toutes les grammaires de la langue s'efforcent de définir la relation de paraphrase au niveau de la phrase, niveau intermédiaire entre celui du texte où se situait l'approche traditionnelle de la reformulation, et celui du mot qu'avaient privilégié les études lexico-sémantiques sur la synonymie. » Fuchs (1994, p. 45)

Le triptyque synonymie, paraphrase, reformulation, dessine donc les contours des études qui s'intéressent à l'équivalence sémantique, aux différents niveaux de l'analyse linguistique.

Les mécanismes de paraphrase qui ont d'abord donné lieu à des descriptions linguistiques systématiques sont limités, car les travaux se bornent généralement aux cas où

l'analyse sémantique est guidée par la prise en compte de critères formels (morphologiques et syntaxiques). Les procédés de paraphrase (recensés par (Dras, 1999)) concernent en particulier les alternances entre voix active et voix passive, les changements de catégorie grammaticale via des relations de dérivation morphologique, des cas d'effacement ou d'insertion, bref pour l'essentiel des « schémas de phrase, à lexique constant » (Fuchs, 1994, p. 72).

Chez Harris, le diagnostic de paraphrase est posé dans le contexte du traitement de l'information : elle est définie comme un principe de préservation du contenu informationnel, une même information pouvant être transmise par des séquences formellement distinctes, obtenues par des opérations de transformation :

> « Alternative relinearizations of a sentence S1 may yield an acceptable paraphrase S2 or, at least, a sentence which is recognizably informationally equivalent to S1. » (Harris *et al.*, 1989, p. 117)

Le sentiment d'équivalence sémantique dépasse néanmoins ces configurations canoniques, et on se rend compte alors que les critères formels que manipule la linguistique ne permettent pas d'apprécier le phénomène dans toute sa complexité – c'est ce que montre Fuchs (1982) en se penchant sur les aspects énonciatifs de la question. D'autres propositions ont élargi le champ d'étude et dégagé le phénomène de paraphrase du cadre restreint de la paraphrase logique auquel il a été longtemps cantonné.

Dans la théorie sens-texte, la notion de paraphrase s'enrichit considérablement. Elle est conçue comme le pivot du modèle :

> « La théorie Sens-Texte se donne, en tant que tâche tout à fait primordiale, une description formelle et exhaustive de la synonymie de phrases, c'est-à-dire de la richesse paraphrastique. » (Mel'čuk, 1988, p. 25)

Le « caractère essentiellement approximatif du lien paraphrastique » est assumé par (Milićević, 2003, 6), dans le même cadre théorique : sa description passe par la définition de rapports de dérivation sémantique qui ne se limitent pas aux cas de relations formelles, morphologiques, mais regroupent des liens sémantiques plus généraux. Ainsi, des cas de paraphrase approximative sont considérés à travers la notion de « (quasi-)synonymie au sens large », qui relie des phrases dont le sens propositionnel n'est pas intégralement superposable. En dernier recours, « seules les données extralinguistiques (connaissance de la situation réelle) peuvent décider si la perte/l'ajout de l'information résultant de la substitution est ou non acceptable » (*op. cit.*, 43). C'est le cas de l'exemple suivant commenté par Milićević :

(1) a. *The victim was stabbed to death.*

b. *The victim was killed with a knife.*

La première phrase précise la nature de l'acte, la deuxième la nature de l'arme. Des règles d'équivalence approximative sont donc prises en compte. La notion de dérivation sémantique est de fait centrale pour le modèle : Mel'čuk et Polguère (2007) désignent par ce terme une « relation entre deux lexies fondée sur une parenté de sens. » Les deux lexies doivent comporter des composantes de sens communes, et la relation qui les unit doit être

récurrente dans la langue, ce qui renvoie à trois grands types de relations entre deux lexies L1 et L2 :

- L1 et L2 ont (à peu près) le même sens : synonymie exacte ou approximative, conversion, dérivation, généricité ;
- L1 et L2 ont un sens opposé ;
- L2 désigne un élément de la situation décrite par L1, typiquement un actant (*tuer / victime*) ou un circonstant (*couper / couteau*).

Enfin, appliquée à l'étude du discours, la notion de paraphrase se dégage totalement de critères formels énoncés *a priori*. Le phénomène de paraphrase discursive, étudié en particulier à travers les procédés de reformulation, s'intéresse à l'équivalence entre unités telle qu'elle est posée par un discours particulier. La relation de paraphrase peut ne pas préexister ni perdurer en dehors de lui. Mortureux (1993) montre ainsi que des syntagmes peuvent fonctionner dans le discours comme des coréférents, quelle que soit leur relation sémantique en langue. Par exemple, les formulations *écart de température* et *déséquilibre thermique* sont mis en équivalence dans le discours par un procédé de juxtaposition, dans l'extrait suivant (cité par M.-F. Mortureux, p. 120) : « Grâce à cet écart de température, à ce déséquilibre thermique... ». M.-F. Mortureux introduit en outre la distinction entre paraphrase *in praesentia* et paraphrase *in absentia,* qui correspondent à deux types de liens construits dans le discours. La première, illustrée par l'exemple précédent, est signalée dans le discours par des procédés de reformulation de nature métalinguistique et typographique. La deuxième « affecte la forme d'une pure et simple substitution » (*op. cit.*, 124) entre deux unités dont l'équivalence n'est pas marquée par un reformulant explicite, mais que le discours regroupe au sein d'un même paradigme, par le jeu des équivalences distributionnelles :

> « [L]a perception [de la paraphrase *in absentia*] repose sur une analyse distributionnelle établissant une équivalence (formelle) entre des syntagmes qui, quelle que soit leur relation sémantique en langue, peuvent fonctionner dans le discours donné comme des coréférents. » (*ibid.*)

Mortureux donne l'exemple des prédicats « être au centre du monde » et « éclairer le monde », dont l'équivalence est repérée parce qu'ils partagent le même sujet « le soleil ».

La paraphrase est donc un phénomène qui met la linguistique en difficulté : si on se contente de critères d'équivalence sémantique stricts, les phénomènes circonscrits sont très limités. Si on s'en écarte, c'est au jugement extralinguistique qu'on se mesure, ou à la spécificité des effets de mise en discours. De son côté, le TAL a placé cette question au premier plan.

1.1.2 La paraphrase reconsidérée par le TAL

Beaucoup d'applications de TAL sont concernées plus ou moins directement par la question de la variabilité linguistique. Le calcul de la proximité sémantique entre des fragments de textes formellement différents est devenu un thème majeur des recherches en TAL. C'est le cas de la recherche d'informations, pour laquelle l'amélioration du rappel des systèmes passe par des procédures d'extension des mots d'une requête ou d'indexation sémantique pour augmenter les chances d'apparier des formulations distinctes d'un contenu

similaire ; c'est aussi le cas des systèmes de questions-réponses qui doivent être capables de repérer des variations entre l'expression de la question et celle de la réponse (Lin et Pantel, 2002). Cette question concerne également les programmes de résumé automatique, puisque l'identification d'informations redondantes permet de rapprocher et de fusionner des passages d'un texte (Barzilay *et al.*, 1999). Enfin, la traduction automatique doit également prendre en compte cette question, par exemple pour rechercher des traductions de mots ou de séquences similaires lorsque la traduction d'un mot donné fait défaut, ou pour mettre en place des procédures d'évaluation (Callison-Burch *et al.*, 2008).

Diversité des segments paraphrastiques considérés On veut donc, dans toutes ces situations, augmenter les chances de détecter des segments textuels sémantiquement proches. Ces segments en relation d'équivalence sont de nature variée. On peut observer que les travaux sur le repérage de la paraphrase se sont peu à peu distanciés du niveau de l'unité lexicale pour aborder le plan syntaxique. La recherche d'information s'est d'abord intéressée à des relations entre index, qui sont le plus souvent des mots isolés, plus rarement des séquences complexes. Des travaux en terminologie ont ensuite étendu l'empan des unités considérées pour étudier le niveau des termes complexes, et en particulier prédire les variantes discursives de termes nominaux (Jacquemin et Tzoukermann, 1999), comme dans l'exemple suivant :

(2) *controllers for signals*→ *signal frequency controllers*

Les composés nominaux (Rosario et Hearst, 2001), ou les combinaisons adjectifs-noms (Lapata, 2001) ont été également étudiés dans le but de calculer des paraphrases verbales de ces unités :

(3) *a migraine treatment*→ *a treatment that prevents migraine*

(4) *an easy problem*→ *a problem that one can solve easily*

Dans les études dédiées plus généralement au repérage de la paraphrase, qui se sont développées au début des années 2000, des patrons morpho-syntaxiques de type varié sont mis au jour. Ainsi, les travaux de Barzilay et McKeown (2001) mettent en correspondance des séquences comme[1] :

(5) $\mathrm{VB}_0\mathrm{VB}^1 \leftrightarrow \mathrm{VB}_0\mathrm{VB}^1$ (*start to talk* ↔ *start talking*)

(6) $\mathrm{IN} : \mathrm{NN}^0 \leftrightarrow \mathrm{VB}^0$ (*in bottles* ↔ *bottled*)

Enfin, Ibrahim *et al.* (2003) ont proposé une méthode permettant de traiter des paraphrases dites structurelles, c'est-à-dire reliant des fragments d'arbres syntaxiques, comme :

(7) *put on*$(\mathrm{X}_{\mathrm{SUJ}},\mathrm{Y}_{\mathrm{OBJ}}) \leftrightarrow$ *wear*$(\mathrm{X}_{\mathrm{SUJ}},\mathrm{Y}_{\mathrm{OBJ}})$

Avec la mise en correspondance d'ilôts prédicatifs, on se rapproche du niveau de la proposition que considèrent les conceptions linguistiques de la paraphrase. En effet, si l'on se réfère à Fuchs (1994), l'utilisation que fait le TAL du terme de paraphrase est abusive, puisqu'elle ne peut être considérée en-deçà du niveau de la phrase :

1. Dans les patrons proposés en exemple, le chiffre en indice marque une identité au niveau du lemme, le chiffre en exposant une identité au niveau de la racine morphologique.

> « [U]ne relation de paraphrase ne peut en effet être posée qu'entre des unités constituant une prédication susceptible de porter une assertion. (...) deux morphèmes ou deux syntagmes synonymes ne sauraient être considérés, en toute rigueur, comme des formulations équivalentes (...) ». (*op. cit.*)

Or, c'est bien toute la variété des points d'ancrage de l'équivalence sémantique que vise le TAL, qui va s'intéresser à des fragments de texte dont l'identité n'est pas toujours claire, mais qui peuvent potentiellement participer au repérage de zones (phrases, passages, voire textes) sémantiquement équivalentes. Les approches dites *data-driven*, basées sur des procédures d'acquisition extensives appliquées à des corpus parallèles ou comparables ou sur le web, fournissent des données qui amènent à diversifier l'empan et la nature des segments candidats.

Nouveaux critères d'évaluation L'inscription de la problématique de la paraphrase dans le TAL amène aussi à définir des critères d'évaluation nouveaux. Dans les contextes applicatifs que nous avons évoqués, décider si deux segments textuels seront considérés comme proches revient à déterminer s'ils apportent un contenu informationnel similaire à l'utilisateur du système qui va les associer. C'est l'approche behavioriste qu'analyse (Condamines, 2000, p. 13) lorsqu'elle dit que « la demande sociale (...) semble plutôt relever du traitement de l'information (...) que du traitement du sens. » Ainsi, la variabilité linguistique peut se mesurer en termes de capacité à mettre en correspondance des questions exprimées par l'usage et des réponses fournies par le système :

> « Dans le cadre de notre système de questions-réponses, nous (...) considérons qu'une expression Y paraphrase une autre expression X si Y permet de répondre à une question contenant l'expression X. » (Duclaye, 2003, p. 16).

C'est donc la tâche qui doit fournir l'instance de validation (on parle alors d'évaluation externe).

Il faut pour l'instant relativiser l'impact de ce changement d'orientation méthodologique : à l'heure actuelle, si quelques rares expériences d'évaluation sont effectivement guidées par la tâche (en traduction automatique en particulier), la majorité des travaux continuent à solliciter le jugement de locuteurs pour réaliser une validation manuelle, l'amenant à se prononcer sur la possibilité que deux formulations soient substituables (*approximately conceptually equivalent*, *roughly interchangeable*). Ibrahim *et al.* (2003) font état de la faiblesse du taux d'accord inter-juges dans la tâche d'évaluation de la paraphrase, au point qu'il s'est avéré impossible d'établir une annotation de référence pour calculer les performances de leur système en termes de rappel. Une proposition de métrique a néanmoins été proposée récemment par Callison-Burch *et al.* (2008).

Malgré cette réserve, les travaux menés en TAL ouvrent donc deux directions pour étendre la réflexion linguistique sur la question de l'équivalence et de la proximité sémantique :

- ils permettent d'envisager des pistes d'évaluation nouvelles, en amenant cette question sur le terrain applicatif ;

– le fait de considérer le besoin informationnel de l'utilisateur d'un système de TAL comme le verdict ultime du jugement de proximité sémantique ouvre très largement le champ des phénomènes que l'on est suceptible de prendre en compte :

> « La réponse à une question est en effet souvent exprimée à l'aide de termes qui ne sont que faiblement (...) liés à ceux de la question. Ainsi l'expression « X a causé Y » peut être considérée comme sémantiquement similaire à « la responsabilité de Y est attribuée à X » dans le contexte des Questions-Réponses (Lin & Pantel, 2001). »(Duclaye *et al.*, 2003)

Dans ce champ de recherche général, je me suis intéressée à la question de la proximité sémantique entre mots. Là tout particulièrement, le jugement de similarité excède les catégories traditionnellement proposées en linguistique par la sémantique lexicale.

1.2 Relations de proximité entre mots : au-delà des relations lexicales classiques

Les relations de paraphrase sont généralement ancrées dans des relations entre mots. Dans ce qui précède, nous avons vu que les deux fragments en relation de paraphrase peuvent comporter les mêmes lemmes (2), des lemmes partageant une même racine morphologique (4), ou des lemmes qu'aucun critère formel ne permet de rapprocher. Dans ce dernier cas de figure, on peut s'attendre à ce qu'une relation de synonymie puisse fournir l'amorce, au niveau lexical, de la relation de paraphrase. C'est le cas de l'exemple (7), qui relie *wear* et *put on.* Mais la relation sémantique entre *treatment* et *prevent* (4), ou entre *problem* et *solve* (3), ne relève pas de la synonymie, puisque les éléments associés n'appartiennent pas à la même catégorie grammaticale. De façon générale, pour estimer la proximité sémantique entre des unités textuelles, le TAL incite à s'intéresser à des relations sémantiques tous azimuts entre mots, sans se limiter nécessairement aux cas mieux circonscrits des relations de synonymie ou de généricité.

Les travaux menés sur la constitution et l'exploitation de thesaurus dans des domaines spécialisés montrent l'intérêt d'intégrer des relations diverses entre mots (Tudhope *et al.*, 2001), faisant appel, au-delà des relations classiques d'équivalence et de hiérarchie, à des relations conceptuelles de nature associative (relation RT : *related terms*). Ainsi, parmi les relations associatives non hiérarchiques définies par Neelameghan (2001) dans le domaine religieux figurent des liens comme *prophet* ABILITY FOR *forecasting*, *cross* MADE OF *wood*, *priest* LOCATION *temple*, *halo* ASSOCIATED WITH *angel*, etc. Kilgarriff et Yallop (2000) distinguent les ressources lexicales qui recensent des relations lâches (*looser lexicons*) ou strictes (*tighter lexicons*), et considèrent que pour certaines applications de TAL, toute relation sémantique peut être bonne à prendre :

> « For Information Retrieval purposes such as finding related documents, connectedness is of interest irrespective of the semantic relation. »(*op. cit.*, p. 1374)

Le spectre des relations sémantiques considérées est très large, et estompe la distinction entre relations syntagmatiques et paradigmatiques, reformulée par Grefenstette (1994a)

en une distinction entre affinités de premier ou de second ordre entre mots. En effet, les relations taxonomiques et de synonymie d'une part, et les relations de nature associative de l'autre, sont toutes susceptibles de pouvoir alimenter les ressources lexicales utilisées par les systèmes d'accès à l'information textuelle (Van der Plas, 2008).

Le principe de relations sémantiques entre mots s'étendant au-delà des relations lexicales habituellement recensées n'est pas sans intérêt, en retour, pour la réflexion linguistique : parallèlement, des travaux sur le discours ont montré ainsi que le phénomène de cohésion discursive est alimenté par des relations entre mots qui ne se limitent pas aux liens de synonymie ou de généricité. Ainsi, Morris et Hirst (2004) affirment la nécessité d'étudier des relations qu'ils qualifient de « non-classiques » (*non-classical lexical semantic relations*), s'inspirant de la distinction posée par Lakoff entre catégories classiques et non classiques.

> « We will extend Lakoff's terminology and refer to relations that depend on the sharing of properties of classical categories as classical relations. Hence we will use the term non-classical for relations that do not depend on the shared properties required of classical relations. » (*op. cit.*, p. 47).

Les auteurs identifient trois principaux types de relations non-classiques étudiés dans la littérature :

- les relations entre mots au sein des catégories non-classiques de Lakoff. Ces relations ne sont pas nommables, sinon par référence à la catégorie dont les mots relèvent lorsqu'on les considère ensemble – ainsi, la référence à l'univers de la bibliothèque permet de relier les mots *inventaire*, *rayonnage*, *prêt*.
- les *case relations* (relations thématiques) : elles associent un prédicat et un de ses actants typiques : agent (*chien* et *aboyer*), objet (*conduire* et *véhicule*), instrument (*couper* et *couteau*), etc.
- les relations étiquetées *related terms* (RT), évoquées plus haut, qui ont été recensées dans le champ des sciences de l'information et de la documentation et relèvent d'une large gamme de relations associatives, comme la localisation, la composition, etc.

S'intéressant à la perception de ces relations non-classiques entre mots dans les textes, Morris et Hirst (2004) montrent que des lecteurs à qui l'on demande de surligner les mots qu'ils perçoivent comme reliés font massivement appel à ces types de relations, les liens de synonymie et d'hyperonymie étant très minoritaires. La description des résultats reste préliminaire, dans la mesure où une typologie reste à élaborer. On retrouve des instances des relations précédemment listées : *case relations*, relations nécessitant la référence à une notion complémentaire commune (*kind* et *gentlemanly* sont reliés par le thème *care*), complétées par des relations de localisation (*funeral* / *chapel*), de cause (*homeless* / *drunk*), etc. Cette expérience conforte la notion de collocation au sens de Halliday et Hasan (1976) – associations typiques entre items lexicaux qui cooccurrent de façon régulière – sur laquelle nous reviendrons dans le chapitre 4 (section 4.2).

L'objectif que je poursuis est précisément de travailler à l'acquisition et l'utilisation de ces relations non classiques, tout en essayant de mieux comprendre leur nature sur le plan linguistique. Le champ considéré est vaste et très peu spécifié puisqu'il n'est considéré que

négativement, comme le montre la qualification de relation « non-classique » que j'utiliserai dans le mémoire à la suite de (Morris et Hirst, 2004). J'ai choisi de me focaliser sur un premier type de relation non classique, à savoir la relation intercatégorielle, et plus particulièrement la relation nom-verbe. C'est de ce type de relation qu'il va d'abord être question dans les deux chapitres suivants, avant que soit considérée, avec l'apport des techniques d'analyse distributionnelle automatique, une façon d'appréhender ces relations non-classiques dans une perspective plus générale.

1.3 Le cas du lien sémantique entre noms et verbes

Mon intérêt pour l'étude de relations sémantiques atypiques entre mots s'est d'abord focalisé sur le cas de la proximité sémantique entre noms et verbes. Cette première direction de recherche trouve son origine dans mon travail de thèse (Fabre, 1996), consacré à l'étude du sens des composés nominaux, qui passe par l'explicitation de la relation prédicative, souvent implicite, qui relie les noms qui forment le composé. Je l'ai déjà illustrée plus haut (3). En voici un autre exemple en français (cf. 2.1) :

(8) *magasin de jouets* →VENDRE(LIEU:*magasin*, OBJET:*jouets*)

La représentation de la sémantique du composé passe par l'identification du prédicat sous-jacent, et l'instanciation de certaines de ses positions argumentales par l'un des éléments du composé, ou les deux. Ce premier travail m'a amené à prolonger l'examen du rapport sémantique entre noms et verbes, à travers plusieurs expériences que je retracerai dans les deux chapitres suivants : repérage de variantes verbales de termes nominaux (section 2.2), construction d'un lexique sémantique inspiré du modèle du Lexique Génératif (section 2.3), calcul de la proximité distributionnelle entre noms et verbes (section 3.4.2). Sachant que les requêtes adressées à des bases de données textuelles sont très massivement formulées à l'aide de noms, j'ai voulu explorer la possibilité d'étendre la formulation initiale à des équivalents verbaux et je me suis intéressée aux conditions du repérage de segments verbaux sémantiquement proches, comme par exemple :

(9) *anniversaire du débarquement / célébrer le débarquement*

(10) *collision de navires / deux navires se sont heurtés*

En dehors du cadre de la théorie Sens-Texte, la linguistique s'est peu intéressée à ces types de liens, qui ne reçoivent pas d'attention particulière dans les travaux consacrés à l'étude des relations lexicales, lorsque le lien sémantique n'est pas accompagné d'une parenté morphologique. En amont de cette perspective applicative, je m'attache donc à dégager et à caractériser cet objet d'étude sur le plan linguistique. Je fais ici le point sur les rares éléments de description linguistique que j'ai pu rassembler pour aborder cette question.

1.3.1 Poser des ponts entre les catégories

Les descriptions qui associent des noms et des verbes sont habituellement, et logiquement, décrites sur le plan syntagmatique. C'est par exemple le cas des fonctions lexicales

qui permettent de décrire les mécanismes de collocation (Wanner, 2004) (par exemple, dans le cas des verbes supports : $Func_0(rumour) = circulate$, $Real_1(promise) = keep$). Ces relations figurent parmi celles que détecte une mesure comme l'information mutuelle, comme l'avaient montré Church et Hanks (1990) au début des études extensives menées sur l'extraction de cooccurrents, en associant analyse syntaxique et mesure statistique : ils montraient que l'on peut par exemple extraire d'un corpus les actions que l'on effectue typiquement avec un téléphone (*answer*, *sit by*, *hang up*, *install*, etc.).

Nous nous intéressons cependant à l'exploration des relations dites de deuxième ordre entre noms et verbes, celles qui s'assimilent à une proximité de nature paradigmatique. Il y a apparemment contradiction dans les termes. Selon (Cruse, 1986) en effet, l'appartenance à une même partie du discours est la condition première d'existence d'une relation lexicale de nature paradigmatique entre deux lexèmes. Dans le cas contraire, Cruse considère que l'on a affaire à des quasi-relations, désignées sous le terme de paronymes, pour lesquelles une relation morphologique vient généralement doubler la relation sémantique (ex : *white/whiten*, *inflate/inflator* : « the meaning of the verb encapsulates the meaning of the adjective » (*op. cit.*, p. 130).

En linguistique théorique et descriptive et en psycholinguistique, des travaux cherchent cependant à décloisonner les catégories, montrant que ces distinctions grammaticales ne sont pas nécessairement opérantes sur le plan fonctionnel. Ainsi, Nespoulous et Virbel (2003) plaident pour l'abandon de la dichotomie nom-verbe de manière à étendre l'observation de relations d'équivalence en discours à des entités que la description grammaticale traditionnelle isole les unes des autres. La porosité de la frontière entre catégories sur le plan cognitif est également attestée par Murphy (2003) :

> « [...] if lexical relations relate word-concepts, rather than lexical entries, the relations can utilize information about words that transcends the lexical boundaries among words. ». (p. 58)

Sur le plan linguistique, (Fellbaum, 1995) donne des arguments empiriques décisifs en montrant à l'aide de méthodes sur corpus que dans le discours, les liens d'opposition qui se tissent entre les mots ne tiennent pas compte des catégories :

> « [r]egardless of their syntactic category, words expressing semantically opposed concepts tend to be used together in the same sentences. » (p. 289)

Elle montre que des paires telles que *begin* (V) et *endless* (Adj), ou *death* (N) et *live* (V) (qu'elle nomme *cross-categorical antonym pairs*), qui marquent une opposition sémantique, présentent également la propriété, typique des paires antonymiques, d'apparaître conjointement de façon significative dans les textes.

Le principal angle d'approche pour aborder les relations de sens entre mots de catégorie différente est celui qu'offre la morphologie : c'est en effet la proximité morphologique qui amène à envisager des passerelles entre les catégories. De fait, on comprend bien dans ce cas que les paires de mots que l'on recense (*manifester/manifestation*, *stocker/stockage*) ne correspondent pas à une relation de nature syntagmatique. Le rôle d'une ressource comme VERBACTION, développée par N. Hathout, est précisément de lister les couples verbe:nom tels que le nom est morphologiquement apparenté au verbe et peut être utilisé pour dénoter

l'action ou l'activité exprimée par le verbe. Parmi les objectifs du projet EuroWordNet figurait l'ajout d'interconnexions entre différentes parties du discours, puisque les relations qui structuraient le lexique dans le WordNet originel ne franchissaient pas la barrière des catégories :

> « From an information retrieval point of view the same information can be coded in an NP or in a sentence. By unifying higher-order nouns and verbs in the same ontology it will be possible to match expressions with very different syntactic structures but comparable content. » (Vossen, 1998)

Ces liens intercatégoriels (*explicit cross-part-of-speech relations*) prennent la forme de relations qualifiées de para-synonymie (*near_ synonymy*) et para-hyponymie, par exemple :

$adorn_V \rightarrow$ XPOS_NEAR_SYNONYM $\rightarrow adornment_V$

Les mots *adorn* et *adornment* sont considérés comme similaires par le sens, même s'ils relèvent de catégories grammaticales distinctes (XPOS). Dans la version actuelle du WordNet de Princeton, on peut utiliser une relation spécifique dénommée *derivationally related form*. Rien n'est donc proposé pour relier noms et verbes au-delà du critère morphologique. On peut néanmoins faire l'hypothèse que la passerelle morphologique permet d'amorcer le calcul de liens à plus longue distance, par exemple le cheminement *adornment*-DERIV-*adorn* + *adorn*-SYN-*decorate* permet de rapprocher *adornment* de *decorate*. Mais d'autres liens intercatégoriels (de type *magasin-vendre*, *livre-écrire*) ne peuvent pas s'appuyer sur un lien morphologique intermédiaire.

1.3.2 Modélisations

L'appartenance de deux mots à deux catégories distinctes semble interdire leur mise en relation sur le plan paradigmatique, puisqu'ils ne sont pas substituables. Mais cette relation peut être elle-même un support pour repérer des entités plus larges (syntagmes, propositions) qui, elles, sont en relation d'équivalence sémantique (typiquement, un GN et une proposition). Cette hypothèse est confortée par les propositions issues de deux cadres théoriques particuliers, le lexique génératif et la théorie Sens-Texte, dont je présente brièvement des éléments auxquels je me réfèrerai par la suite.

1.3.2.1 Dans la théorie du lexique génératif

Le lexique génératif définit le principe d'une structure prédicative et événementielle attachée aux unités lexicales, permettant ainsi de proposer un modèle explicatif global de la polysémie. On peut partir, comme le fait Pustejovsky (1995), du constat que des contextes linguistiques manifestent un type particulier de polysémie nominale, et amènent à interpréter un nom concret comme un événement :

(11) a. *J'ai commencé le livre que tu m'as donné.*

b. *C'est un livre difficile, inclassable mais jubilatoire.*

c. *Difficile de partir en vacances sans prendre un livre de plage avec soi.*

d. *Avant ce livre*[2]

Ces quatre exemples illustrent quatre types de contextes qui déclenchent une interprétation événementielle du nom *livre* : la position objet du verbe *commencer*, la modification par l'adjectif *difficile*, la composition nominale (de structure *N N* en anglais et *N prép N* en français), l'apparition dans une structure *avant* + SN. Ce type d'exemples montre qu'un nom comme *livre* est susceptible de projeter une information événementielle spécifique (l'événement *lire*) lorsque celle-ci est déclenchée dans des contextes particuliers. Selon J. Pustejovsky, le mécanisme de coercion n'est possible que parce que cette dimension événementielle est déjà intégrée dans la sémantique du nom. Il propose donc une représentation sémantique riche des unités lexicales, qui comporte cette information événementielle, permettant ainsi de modéliser ces mécanismes de coercion et d'expliquer de manière régulière ces cas de polysémie.

Toutes les catégories syntaxiques reçoivent ainsi le même type de représentation. Une unité lexicale (UL) est munie d'une structure argumentale et d'une structure événementielle, qui indique le type d'événement dénoté par l'unité (*states*, *processes*, *events*, etc.). La structure des *qualia* détermine quatre attributs essentiels du sens des mots, et structure les informations associées à l'UL en liant les arguments et les événements entre eux. Le rôle formel (FORMAL) associe à l'UL sa classe sémantique, ce qui correspond pour *livre* aux deux types INFORMATION et OBJET_PHYSIQUE. Le rôle constitutif (CONST) définit la relation entre l'UL et ses composantes (*pages*, *couverture*, *texte* pour *livre*). Le rôle télique (TELIC) identifie la finalité de l'UL (*lire*, *publier*, etc.). Le rôle agentif (AGENT) spécifie les conditions d'existence (le livre peut être écrit ou fabriqué).

Les liens sémantiques entre noms et verbes sont donc formulés dans le lexique génératif par le biais de rôles abstraits qui sont instanciés dans des contextes particuliers. J'ai fait appel à ce modèle à plusieurs reprises : pour expliquer le fonctionnement de la composition nominale dans le cadre de ma thèse (cf. 2.1), pour guider une procédure d'acquisition de liens nom-verbe (cf. 2.3), et pour étudier l'application du phénomène de coercion au contexte événementiel *avant* + GN, dans un travail mené avec Anne Le Draoulec (Le Draoulec et Fabre, 2006).

1.3.2.2 Dans la théorie Sens-Texte

Dans la théorie Sens-Texte, les Fonctions Lexicales (FL) permettant de rendre compte de la dérivation sémantique (1.1.1) sont considérées comme des FL paradigmatiques par opposition aux FL syntagmatiques chargées de décrire les phénomènes de collocation. Certaines concernent les relations de synonymie (Syn) et d'antonymie (Anto). Les autres établissent des liens entre deux lexèmes relevant de catégorie différente, du verbe vers le nom (S_0, S_i), ou du nom vers le verbe (V_0). La FL S_0 relie un verbe et son correspondant nominal ($S_0(acheter)$=*achat*, $S_0(dormir)$=*sommeil*). La FL V_0 relie un nom et son correspondant verbal ($V_0(promesse)$=*promettre*). Les FL S_1,S_2, S_3... désignent le 1^{er}, 2^e, 3^e actant du

2. A distinguer de l'interprétation existentielle : *les scandales sont apparus bien avant ce livre* = avant qu'il n'existe. Cf. Le Draoulec et Fabre (2006)

mot clé (S_1(*écrire*)=*auteur*, S_2(*écrire*)=*livre*)). On comprend mieux la dimension paradigmatique de ces FL lorsqu'on observe leur contribution à des règles de paraphrase, par exemple dans les « règles de fission » suivantes (Mel'čuk, 1988, 39) :

$S(C_0) \leftarrow Oper_2(S_0(C_0))$

(*Jean nous a bien accueillis*$[C_0] \Longleftrightarrow$*Nous avons trouvé* $[Oper_2(S_0(C_0))]$ *chez Jean un bon accueil* $[S_0(C_0)]$)

$S(C_1) \leftarrow Oper_1(S_1(C_1))$

(*Belfort, qui a écrit* $[C_0]$ *cet article* $\Longleftrightarrow$*Belfort, qui est* $[Oper_1(S_1(C_1))]$ *l'auteur* $[S_1(C_1)]$*de cet article*

La première règle est basée sur la substitution au prédicat verbal *accueillir* (X *accueillir* Y) d'une structure à verbe support de type Oper_2 qui s'appuie sur la forme nominalisée du verbe ($S_0(C_0)$) (Y *trouver un accueil chez* X). La deuxième est basée sur la mise en correspondance entre le verbe *écrire* et son premier actant *auteur* ($S_1(C_1)$).

Ces deux exemples montrent que le phénomène de paraphrase combine des règles de dérivation sémantique (lien entre un verbe et le nom d'action correspondant, ou entre un verbe et le premier actant) et des contraintes syntaxiques qui contrôlent l'application de ces règles lexicales. On voit ainsi que les deux familles de FL (syntagmatiques et paradigmatiques) ne sont pas étanches, dans la mesure où les relations paradigmatiques (ici, de type S_0 ou S_i) peuvent être incorporées dans un processus de réagencement syntagmatique (par l'utilisation de fonctions de type $Oper_i$) .

1.4 Conclusion

Les approches TAL caractérisées par une démarche *data-driven* et, au moins en théorie, *task-driven* offrent un cadre d'analyse qui renouvelle doublement les objets que l'on considère. Tout d'abord, la tâche dicte de nouveaux objectifs et de nouveaux modes d'évaluation : on s'intéresse à la proximité sémantique dès lors que deux segments textuels peuvent répondre à un même besoin informationnel. Ensuite, le recours à des techniques de corpus extensives – aligner des traductions, rechercher les instances d'un patron paraphrastique sur le web – produit des données d'observation qui dépassent les limites des schémas paraphrastiques déjà établis.

La notion de paraphrase telle qu'elle est définie en linguistique ne couvre alors qu'un sous-ensemble des phénomènes à traiter, et il semble important que la réflexion linguistique travaille à la spécification des objets langagiers que le TAL contribue à faire émerger. Cette perspective m'a amenée à repérer dans les travaux de linguistique des pistes de travail qui me semblaient mériter d'être développées, pour chercher à étendre le repérage de l'équivalence sémantique au cas des relations non classiques, et plus particulièrement au cas des relations intercatégorielles comme type d'ancrage particulier de la relation d'équivalence. Cette direction de recherche amène à modifier le point de vue sur certains objets linguistiques, par exemple dans cette réflexion sur la non-étanchéité des dimensions syntagmatique et paradigmatique. Le travail sur les relations intercatégorielles brouille en effet

cette frontière. Ainsi de la relation entre *livre* et *lire* : en tant qu'objet du verbe, on peut le trouver dans une situation de cooccurrence :

(12) a. *Lire des livres prolonge l'existence.*

b. *Voilà un livre à lire lentement.*

Mais le nom peut également apparaître, sans le verbe, dans un contexte où il incorpore l'information prédicative et la rend disponible :

(13) a. *Avant ce livre je ne connaissais pas cet auteur.*

b. *Commencez par un livre facile.*

Dans le chapitre suivant, je présente deux expériences sur ce sujet, qui visent l'une le repérage de paraphrase (ou de variation), l'autre la constitution de ressources. Toutes les deux concernent le lien nom-verbe, mais la première est basée sur la présence d'un lien morphologique (ex : *comparaison de résultats* / *comparer des résultats*), l'autre au contraire cherche à s'en affranchir en repérant des liens téliques (ex : *livre* / *lire*, *prise* / *brancher*).

Chapitre 2

Nouveaux liens de proximité sémantique : rapprochements intercatégoriels

Les deux expériences que je relate m'ont amenée à étudier les liens intercatégoriels entre nom et verbe selon deux points de vue complémentaires, qui mettent en jeu l'un et l'autre la question de l'équivalence ou de la proximité sémantique : la première expérience, menée en collaboration avec Christian Jacquemin, se basait sur l'existence d'un lien morphologique pour repérer des variantes verbales de termes nominaux ; la seconde est le fruit d'un projet collectif de plus grande ampleur, coordonné par Pascale Sébillot, qui visait à construire un lexique de conception nouvelle pour identifier des liens sémantiques entre noms et verbes à partir de techniques d'apprentissage en corpus.

J'amorce dans ce chapitre la présentation des principaux éléments du dispositif dans lequel je travaille, et avec lequel j'ai commencé à me familiariser à l'occasion de ces deux expériences :

- le travail sur corpus est le point commun de tous mes travaux depuis la thèse, et je découvre progressivement les techniques qui permettent d'en faire émerger des informations nouvelles susceptibles de renouveler la description de l'objet étudié. Les travaux présentés dans ce chapitre m'amènent à saisir l'articulation entre les démarches *corpus-based* et *corpus-driven* sur lesquelles je ferai le point dans le dernier chapitre (section 9.2.2).
- le recours à des procédures de TAL variées, faisant appel à différents niveaux de marquage des textes, à plusieurs techniques d'exploration des corpus (grammaires locales, procédures d'apprentissage), et à des critères d'évaluation qui s'affinent peu à peu.

Au-delà de la présentation de l'objet précis de ces travaux, dont j'ai présenté le cadre général dans le chapitre précédent, je commence donc ici une réflexion que j'approfondirai dans la troisième partie du mémoire, concernant le rôle de l'analyse linguistique dans ce type de recherche, et l'apport de techniques de TAL pour approfondir l'examen des notions

qui m'intéressent.

Les travaux que je présente ici prennent tous les deux leur origine dans mon travail de thèse, bien qu'ils relèvent d'hypothèses méthodologiques différentes par rapport à celles que j'avais initialement adoptées. Je commencerai donc par expliquer la continuité entre mon projet de thèse et les expériences qui ont suivi.

2.1 Point de départ : une thèse sur la sémantique des composés nominaux

Ma thèse (Fabre, 1996), encadrée par Pascale Sébillot [1], portait sur la modélisation de la sémantique des composés nominaux en anglais et en français. Elle visait à prédire le lien sémantique, souvent implicite, qui unit les deux noms d'un composé, et à produire la représentation prédicative correspondante, comme dans l'exemple suivant :

(1) *shark killer* → KILL(SUJ:*shark*)

(2) *magasin de jouets* → VENDRE(LOC:*magasin*,OBJ:*jouets*)

Cette thèse est très influencée par les hypothèses de l'intelligence artificielle : elle vise à modéliser la sémantique sous-jacente de ces unités complexes. La tâche s'avère difficile. En effet, des principes structurels tiennent une petite partie de l'édifice : dans le cas des composés à tête déverbale, comme dans l'exemple 1, l'information verbale est accessible et le cadre de sous-catégorisation disponible pour élucider les relations entre les composants, lorsque ceux-ci sont de nature argumentale. Pour le reste, il est difficile de concevoir des principes organisateurs de ces stuctures puisque toute relation pragmatiquement adéquate peut potentiellement connecter deux noms – en anglais tout particulièrement. J'ai trouvé des clés pour aborder ce problème dans le modèle du lexique génératif (Pustejovsky, 1995) (cf. 1.3.2.1). Les composés nominaux sont en effet un cas particulièrement intéressant d'illustration des mécanismes d'incorporation et de déclenchement d'informations prédicatives sous-jacentes dans la représentation de la sémantique des noms. J'ai proposé alors de faire appel à des principes de fonctionnement généraux, basés sur la définition de la structure *qualia* associée aux noms, qui, en étant activée en contexte, donne accès à des *slots* de différents types.

L'objectif de la thèse était également d'utiliser l'information sémantique explicitée pour améliorer les performances d'un système de recherche d'information, de façon à tester l'hypothèse selon laquelle le prédicat verbal qui est calculé peut permettre d'étendre la recherche pour trouver des segments de textes apparentés sémantiquement. Je reviens sur ce point et ses prolongements ultérieurs section 3.5.

J'ai conservé de ce travail de thèse quelques convictions positives et négatives. En positif, j'ai considéré qu'il fallait prolonger l'examen d'objets similaires, en poursuivant deux idées initiées pendant la thèse : le nom active en contexte des informations sémantiques de

1. Thèse de doctorat menée au sein du laboratoire IRISA, Université de Rennes 1. Elle a bénéficié d'un financement INRIA-CNET.

nature relationnelle, qui doivent amener à étudier le lien entre noms et verbes ; ces informations relationnelles étendues peuvent ensuite s'avérer précieuses pour repérer dans une perspective applicative des possibilités nouvelles de rapprochements sémantiques.

En négatif, j'ai acquis la conviction qu'il fallait s'y prendre tout autrement : le traitement par règles a des limites évidentes. Ce type d'information lexicale n'est pas accessible *ex abstracto* mais est lié au type de discours et au contexte dans lequel le nom apparaît. J'ai donc renoncé à l'approche initiale, pour me tourner vers l'analyse de corpus afin d'y repérer des associations sémantiques effectives [2]. Je n'ai pas su prendre pendant la thèse le tournant de la linguistique de corpus. Ce fut chose faite l'année suivante, en 1997, grâce à la collaboration avec B. Habert, qui a débouché en particulier sur la rédaction d'un manuel avec Fabrice Issac (Habert *et al.*, 1997). Les expériences que je présente dans ce chapitre relèvent donc naturellement de cette approche sur corpus.

2.2 Repérage de variantes dérivationnelles de termes

Les termes d'un domaine subissent des variations linguistiques importantes au gré de leurs occurrences dans les textes et il est crucial, dans les traitements basés sur la reconnaissance terminologique, de repérer les termes sous leurs différents modes d'apparition (Jacquemin et Tzoukermann, 1999). Le but de l'expérience que j'ai menée avec C. Jacquemin était de constituer une grammaire permettant d'identifier les variantes verbales de termes nominaux. Il s'agissait de contrôler la proximité sémantique des deux formulations, de manière à récolter comme variantes les séquences verbales jugées sémantiquement équivalentes au terme nominal, et à rejeter celles qui ne présentaient pas de proximité sémantique avec le terme de départ. Ainsi, les séquences suivantes sont considérées comme des couples terme / variante :

(3) *comparaison de résultat / comparer les résultats*

alors que dans l'exemple suivant, la séquence verbale n'est pas une variante du terme nominal et doit être rejetée.

(4) *amélioration d'une technique / techniques permettant d'améliorer*

Nous nous sommes intéressés uniquement à des cas de variation dérivationnelle, pour lesquels nous disposions d'une information relative au lien morphologique entre le nom et le verbe. Ma contribution a donc consisté à étendre aux cas de variation intercatégorielle les règles de variation prises en charge par l'analyseur.

2.2.1 Faster, un outil de repérage de la variation terminologique

Faster est un analyseur syntaxique développé par C. Jacquemin dédié au repérage de la variation terminologique (Jacquemin, 1997). Il repose à l'origine sur les principes suivants :

2. Cette approche sur corpus a donné lieu à des développements intéressants sur les composés en anglais par (Rosario et Hearst, 2001).

- il est dédié au repérage de la variation morpho-syntaxique : ce type de variation n'affecte pas les mots pleins, sinon sur la base de transformations morphologiques[3]. L'analyseur détecte donc des variations élémentaires comme la permutation, l'insertion, la coordination, l'élision, et des variations issues de la composition de ces variations élémentaires ;
- il prend comme point de départ une liste de termes contrôlés, qui sont transformés automatiquement en une structure syntaxique, par « recyclage linguistique », selon les termes de C. Jacquemin ;
- une métagrammaire est conçue pour décrire les variations qui peuvent affecter ces termes initiaux. Les métarègles prédisent les transformations licites dans la structure des termes.

La phase de conception des métarègles est manuelle. Elle est basée sur la connaissance des types de variation possibles, et sur une phase d'« ajustement expérimental de ces données par confrontation aux occurrences observées » (*op. cit.*, p. 88). Des contraintes sont ajoutées aux règles génériques initiales pour filtrer les variantes impropres. Voici un exemple de métarègle détectant un cas de coordination. Le schéma de la règle initiale est donné par la partie gauche de la métarègle, le schéma de la variation par la partie droite.

Métarègle $Coor(N_1 \rightarrow N_2 A_3) = N_1 \rightarrow N_2 A_4 C_5 A_6$
(A_3 lemme) = (A_6 lemme)
(A_3 flexion) = (A_6 flexion)
(A_4 accord genre) = (N_2 accord genre) = (A_6 accord genre)
(A_4 accord nombre) = (A_6 accord nombre)

La partie gauche de la règle est unifiée avec le terme initial (préalablement recyclé en règle). Les traits associés au nom N_2 sont transmis dans la partie droite de la métarègle en raison de la présence du symbole N_2 dans cette partie. Des traits de A_3, seuls les traits de lemme et de flexion sont transmis (il n'est pas nécessaire que A_6 et A_3 soient de même nombre). Cette métarègle permet par exemple de reconnaître, à partir du terme initial *artère cérébrale* la variante *artères carotides et cérébrales.*

2.2.2 Traiter la variation verbo-nominale

Notre objectif était donc d'étendre les patrons de variation que FASTER était à même de repérer, et d'étudier par ce biais les possibilités d'acquisition de terminologie de nature verbale. Les systèmes d'acquisition terminologique se sont en effet principalement concentrés sur la recherche de termes nominaux, défendant l'idée selon laquelle la dénomination et la construction de taxonomies est réalisée essentiellement à l'aide de noms et de leurs expansions. Dans l'analyseur LEXTER (Bourigault, 1994), les formes verbales étaient ainsi considérées comme des marques de frontière utilisables pour délimiter les candidats termes. Il apparaît dans un deuxième temps qu'on ne peut pas exclure ainsi les relations verbales de la description terminologique, dans la mesure où elles contribuent elles aussi à structurer

3. Le traitement de la variation sémantique affectant les mots pleins a été également mis en œuvre par C. Jacquemin, en particulier en collaboration avec Emmanuel Morin puis dans le cadre du développement d'un système de questions-réponses au LIMSI (Ferret *et al.*, 2001).

la terminologie d'un domaine. Selon L'Homme (2002), l'information verbale doit intégrer les dictionnaires spécialisés soit pour recevoir une information pour elle-même, soit pour compléter l'information disponible sur les termes nominaux (Claveau et L'Homme, 2004). Notre perspective relevait du deuxième cas de figure, et consistait plus particulièrement à considérer les éléments verbaux au moment du repérage des termes. Si le groupe *analyse de croissance* est répertorié comme un terme et peut donc être utilisé pour indexer des contenus, il est utile de savoir qu'il peut se retrouver sous une forme verbale dans le groupe *analyser la croissance* ; de même pour :

(5) *calcul de coefficient / coefficient a été calculé*

(6) *variabilité de rendement / rendements variant*

Nous avons proposé une méthode pour accéder à des unités verbales, sans toutefois nous prononcer dans le cadre de ce travail sur leur statut terminologique. Plutôt que d'envisager l'acquisition autonome de groupes verbaux, nous nous sommes appuyés sur un ensemble de termes nominaux déjà recensés pour l'étendre du côté du vocabulaire verbal en utilisant des principes de variation dérivationnelle.

2.2.3 Approche du problème

En prenant en compte le phénomène de la variation dérivationnelle, nous avons élargi la portée des règles de découverte de la variation terminologique. Les variantes que nous avons étudiées sont couvertes par plusieurs catégories de métarègles. Toutes relèvent du cas où les deux groupes (le terme et sa variante) sont liés par une relation de dérivation nomino-verbale. Le nom tête ou le nom modifieur du terme de départ est morphologiquement apparenté à la forme verbale de la variante. Par exemple :

(7) *injection de la solution/ injecter une solution*

L'expérience menée a consisté à définir les contraintes qui permettent de considérer la séquence comme une variante sémantiquement valide du terme nominal attesté. Les patrons syntaxiques ayant permis de récolter ces associations terme / variante ont été mis au point en deux temps. Une première étape a consisté à projeter des métarègles très permissives conçues par Jacquemin *et al.* (1997) pour repérer les associations potentielles entre un groupe nominal et un groupe verbal. La mise au point expérimentale a permis ensuite d'affiner les transformations initiales par l'observation des occurrences trouvées en corpus, et d'améliorer ainsi la précision des patrons de recherche par la définition de nouvelles règles et l'ajout de contraintes.

2.2.3.1 Métarègles génériques

Nos résultats sont basés sur l'étude de 1050 associations. Le terme est un groupe nominal de structure $N_1 P_2 N_3$, avec $P_2 = de$ dans 92% des cas. La variante est un groupe contenant une forme verbale, dans lequel on retrouve les deux mots pleins du terme, soit directement, soit sous une forme dérivée : dans ce dernier cas, le verbe de la variante est morphologiquement lié à un nom du terme. Le tableau 2.1 donne une présentation très

simplifiée des métarègles, dans lesquelles les parenthèses renvoient à des séquences licites de catégories que je ne reproduis pas ici, et le codage $M(N_1)_V$ signifie que l'unité est un verbe et qu'il partage avec N_1 une même base morphologique.

Métarègle	Exemple
$NToV_1(N_1P_2N_3) \rightarrow M(N_1)_V(...)N_3$	*caractérisation de l'activité → caractérise l'activité*
$NToV_2(N_1P_2N_3) \rightarrow N_3(...)M(N_1)_V$	*application de l'analyse → analyse est appliquée*
$NToV_3(N_1P_2N_3) \rightarrow N_1(...)M(N_3)_V$	*méthode de caractérisation → méthode pour caractériser*
$NToV_4(N_1P_2N_3) \rightarrow M(N_3)_V(...)N_1$	*critère d'évaluation → évalués selon les critères*

TABLE *2.1 – Métarègles génériques*

Ces quatre métarègles génériques permettent de prendre en compte les configurations suivantes :

- le déverbal apparaît en tête ou en expansion du terme initial ;
- l'ordre des mots dans la variante est conforme à celui du terme initial (à la dérivation près), ou bien les mots sont inversés.

Ainsi, les métarègles 1 et 3 produisent des groupes qui maintiennent l'ordre des deux mots, alors que les métarègles 2 et 4 les inversent. Dans le cas de 1 et 4, la tête du groupe résultant est le verbe (on obtient un groupe verbal), alors que la forme verbale est en position finale dans les deux autres cas.

La transformation syntaxique est décrite dans l'espoir que les variantes reconnues vont préserver le contenu informationnel du terme. Tout le problème de notre approche a donc consisté à définir des liens de transformations syntaxiques entre une séquence nominale et une séquence verbale qui puissent être rapprochées sur le plan informationnel. Or, les règles ne permettent pas toujours de repérer des variantes correctes du terme. Voici des exemples complémentaires qui illustrent des cas de repérage d'occurrences impropres du terme (l'astérisque marque le fait que les deux séquences ne sont pas dans un rapport de variation valide) :

(8) $NToV_1$ *mutation du gène* $^* \rightarrow$ *mutées dans les gènes*

(9) $NToV_2$ *détermination de sensibilité* $^* \rightarrow$ *sensibilités des glossines saines déterminent*

(10) $NToV_3$ *système de production* $^* \rightarrow$ *systèmes enzymatiques produisant*

(11) $NToV_4$ *processus d'induction* $^* \rightarrow$ *induit des processus*

On peut dire dans ces cas que les spécifications de type morpho-syntaxique n'ont pas suffi à repérer une variante valide. Notre objectif était donc de contrôler l'application de ces règles pour obtenir des séquences associées aux termes de départ, et pour caractériser la relation sémantique qui est induite. Cela revient à éliminer le bruit produit par des règles morpho-syntaxiques trop permissives, pour leur substituer des règles valides sémantiquement.

2.2.3.2 Examen des variantes

On a affaire à un problème de variation bien circonscrit et qui pourtant montre déjà le problème dans toute sa complexité. L'analyse des séquences détectées par l'application des métarègles permet d'identifier deux cas de variantes correctes : le premier correspond à celles pour lesquelles on peut parler d'équivalence notionnelle (tous deux peuvent être utilisés pour renvoyer à une situation similaire) :

(12) *pilotage d'irrigation/ piloter les irrigations*

(13) *air de séchage/ séchées à l'air*

Le deuxième cas de figure concerne des variantes qui correspondent à des exemplaires plus spécifiques du concept de départ : l'ajout d'un adverbe, d'un adjectif ou d'un complément du nom a pour effet de créer une variante hyponyme (à la modalisation près, que peut intégrer la variante verbale, alors qu'elle est en dehors du champ du terme nominal, comme on le voit dans le cas de l'exemple (14)) :

(14) *utilisation de spectrométrie/ spectrométrie d'image pourrait être utilisée*

(15) *application de la méthode/ méthode de sorption oscillante appliquée*

La séquence repérée peut également présenter des variantes plus subtiles du terme de départ, certaines considérées comme négligeables du point de vue terminologique, comme dans le cas de variations de déterminant (*exploitation de forêt / exploiter cette forêt*), d'autres plus difficiles à apprécier. Il est de fait parfois impossible de décider si deux séquences peuvent être ramenées à un même contenu informationnel sans tenir compte du contexte ni faire intervenir des connaissances propres au domaine. Comme le signale Ibekwe-SanJuan (2005), il est difficile d'apprécier la proximité sémantique autrement que de façon intuitive, les techniques classiques de calcul de distance (distance d'édition en particulier) s'avérant difficiles à échelonner. L'effet d'un changement de préposition par exemple (*plantation en sol / plantés sur sol*) est délicat à apprécier sans connaissance experte et ne sera pas le même selon le terme d'origine. Autre exemple, la métarègle 3 ($NToV_3$ dans le tableau 2.1) produit quatre instances qui présentent la même configuration en surface : $N_1P_2DN_4M(N_3)_V$. Dans cette séquence, $M(N_3)_V$ est à chaque fois une forme de participe passé. S'agit-il de variantes correctes ? On voit qu'un schéma identique peut produire des variantes de qualité assez différente :

(16) *densité de inoculation/ densité de la souche inoculée*

(17) *étude de consommation/ étude de la puissance consommée*

(18) *méthode de dosage/ méthode des ajouts dosés*

(19) *qualité de production/ qualité du bois produit*

Si l'on considére que *qualité du bois produit* peut constituer une variante hyponyme de *qualité de production*, on ne peut par contre pas rapprocher *densité d'inoculation* et *densité de la souche parasite inoculée* – il s'agit de deux mesures différentes de la densité. Dans ce cas il paraît difficile de pouvoir généraliser ces observations pour aboutir à une règle capable de réaliser un filtrage adéquat et il apparaît, d'emblée, qu'une approche systématique à base de règles ne permettra pas de traiter des cas de cette nature.

2.2.3.3 Principe de préservation de la structure syntaxique

L'option que nous avons prise a été de traduire cette question en termes sémantiques, c'est-à-dire d'essayer de prendre en compte des caractéristiques qui, en amont de ces aspects pragmatiques et énonciatifs, à partir de principes relevant du fonctionnement linguistique de ces unités, expliquent au mieux la répartition entre variantes et non-variantes. Le critère qui nous a guidés a consisté à distinguer les variantes qui préservent le rapport syntaxique entre les unités du terme initial et celles qui le perturbent. Ce critère permet de différencier les variantes incorrectes et les variantes potentiellement valides, sachant que cette deuxième catégorie est hétérogène et doit être affinée dans un deuxième temps par un avis expert en tenant compte des nuances que nous venons d'apporter. Nous avons donc maintenu le principe d'une approche « permissive » des règles de variation, qui consiste à privilégier le rappel et à s'en remettre ensuite aux effets de filtre du corpus.

La préservation de la structure argumentale est donc le paramètre principal du réexamen des règles de variation que nous avons réalisé. Nous avons proposé une représentation prédicative du syntagme, dans laquelle le nom déverbal ou le verbe sont représentés munis d'une structure argumentale remplie par les arguments indicés 0 (argument sujet ou externe), 1^{er} (1er argument ou argument interne 1), 2^{e}(2ème argument ou argument interne 2) ou *ajout* (argument non essentiel, ajout).

Il existe trois cas de préservation de la structure argumentale :

- Le dépendant nominal est un argument interne de la tête du terme nominal ainsi que du verbe de la variante. Il peut s'agir d'un argument direct ou d'un argument oblique :
  $N(arg_1) \rightarrow V(arg_1)$
  évaluation de l'efficacité / évaluer l'efficacité $(NToV_1)$
  application de l'analyse / analyse est appliquée $(NToV_2)$
  $N(arg_2) \rightarrow V(arg_2)$ *application à l'étude/ appliquées à l'étude* $(NToV_1)$
- Le dépendant nominal est l'argument externe (sujet) de la tête du terme nominal ainsi que du verbe de la variante.
  $N(arg_0) \rightarrow V(arg_0)$
  variation de la production / production varie $(NToV_2)$
  fonctionnement de l'enzyme / enzyme pouvait fonctionner $(NToV_4)$
- Le dépendant nominal joue le rôle d'un argument non thématique, de type circonstant. Celui-ci joue le même rôle sémantique R par rapport à la tête du terme nominal et par rapport au verbe de la variante.
  $N(ajout_R) \rightarrow V(ajout_R)$
  traitement à la chaleur / traitée à la chaleur (moyen) $(NToV_1)$
  acclimatation en serre / acclimatées en serre (lieu) $(NToV_1)$

Les deux premières catégories s'enrichissent d'un cas particulier : dans le terme, la relation argumentale peut être sous-spécifiée et renvoyer à l'argument interne ou externe. Cette ambiguïté n'est pas maintenue dans la forme verbale de la variante. Le dépendant nominal est alors réalisé soit comme objet, soit comme sujet du verbe :

- $N(arg_0|arg_1) \rightarrow V(arg_1)$

variation du temps / varier le temps ($NTo V_1$)
augmentation de concentration / augmente la concentration($NTo V_1$)
- $N(arg_0|arg_1) \rightarrow V(arg_0)$
augmentation de l'intensité / intensité augmente ($NTo V_2$)
arrêt de croissance / croissance s'arrête ($NTo V_2$)

Les cas de distribution argumentale non respectée constituent les configurations complémentaires de celles que nous venons de présenter.

- $N(arg_1)* \rightarrow V(arg_0)$: *détermination de facteur / facteurs déterminent*
- $N(arg_1)* \rightarrow V(arg_2)$: *application de l'analyse / appliqué à l'analyse*
- $N(arg_1)* \rightarrow V(ajout)$: *réalisation du modèle / réalisée sur modèle*
- $N(ajout)* \rightarrow V(arg_0)$: *taux d'augmentation / taux augmente*
- $N(ajout)* \rightarrow V(arg_1)$: *méthode d'utilisation / utilise une méthode*

2.2.3.4 Capter des informations sémantiques à l'aide de marques légères

La dernière étape du travail a consisté à enrichir les métarègles existantes de spécifications supplémentaires pour pouvoir traiter ce niveau de différenciation. En principe, les contraintes qu'il s'agit de contrôler font appel à des connaissances linguistiques riches : pour identifier quel schéma syntactico-sémantique est instancié par une variation, il faut vérifier un ensemble de paramètres qui sont la satisfaction des contraintes argumentales (nombre d'arguments), le respect des restrictions sélectionnelles (type des arguments), la conformité de la sémantique de la préposition avec la valeur de complément considéré. Dans la mesure où il n'était pas possible de se baser sur un lexique assez riche pour réaliser ces vérifications, j'ai proposé des règles reposant sur une description lexicale minimale. J'ai été amenée à ajouter des contraintes supplémentaires dans les règles (concernant le lemme de la préposition, la forme et la voix du verbe), dont certaines sont basées sur la définition de nouveaux traits à intégrer au niveau du lexique, à savoir la nature morphologique du nom déverbal (agentif vs processif) et la valence du verbe. Il était crucial que le processus d'enrichissement du lexique soit peu coûteux. De fait, nous avons estimé à 3 heures le temps nécessaire pour annoter plus de 1000 verbes avec ces nouveaux traits.

Les valeurs possibles du trait de valence sont intransitif, transitif, ditransitif, ergatif. Ce dernier trait permet d'identifier une catégorie de verbes qui acceptent des alternations de type transitif/intransitif : le même SN peut être objet de l'emploi transitif et sujet de l'emploi intransitif. Ce type d'alternance concerne des verbes qui peuvent décrire une action du point de vue du patient ou du point de vue de l'agent : *la température augmente / on augmente la température*. Dans le corpus étudié, ces alternances sont fréquentes et concernent principalement des verbes de changement : *augmenter*, *fermenter*, *muter*, *saturer*, *cristalliser*, etc. La forme intransitive est plus souvent une forme pronominale : *se dégrader*, *s'arrêter*, *se régénérer*, *se transformer*, *se réguler*, etc. La prise en compte de cette caractéristique permet de traiter des cas d'ambiguïté. Parmi les métarègles redéfinies à l'issue de ce travail, figure par exemple la suivante, qui prend en compte ce critère de valence verbale pour les variantes du type *chute de température / température chute* :

$NtoV(N_1 P_2 N_3) \rightarrow N_3\,(...)\,M(N_1)V$

$P_2 = de(N_1 dev) = processive \wedge (M(N_1)V\,tense) \neq pastparticiple \wedge (M(N_1)V\,valence) = ergative \mid intransitive$

La contrainte sur la valence (*ergative* ou *intransitive*) permet de bloquer des équivalences du type *recherche de programme / programme recherche*.

Je n'ai ici donné qu'une version très simplifiée des métarègles, qui sont décrites dans le détail dans (Fabre, 1998) ; les résultats et leur évaluation ont été présentés dans (Fabre et Jacquemin, 2000) : nous avons montré que les performances (en termes de rapport rappel / précision) sont nettement améliorées à la fois par rapport à une *baseline* consistant à extraire des segments où les mots pleins du terme initial cooccurrent, et par rapport à l'application des métarègles génériques que nous avons enrichies.

2.2.4 Bilan de cette expérience

Sur le plan méthodologique, ce travail montre une modalité de l'intervention linguistique à travers l'intégration d'informations dans un système qui n'est pas conçu pour supporter des descriptions lexicales fines qui supposeraient un travail considérable de mise à jour pour s'adapter à la terminologie d'un domaine d'activité spécifique. Dans ce type de tâche, le linguiste doit admettre que le verdict ultime de son travail est la performance de l'analyseur vis-à-vis de la tâche considérée – un compromis entre rappel et précision. Et il doit mener une double démarche : tout d'abord, il doit traduire le problème en termes de contraintes linguistiques. On l'a vu, la règle linguistique proposée n'est qu'une approximation du diagnostic de variation. Le traitement linguistique doit être à même de réduire le problème pour qu'il puisse être traité. L'observation des données montre que la préservation du rapport argumental entre les composants des deux séquences est une condition nécessaire mais pas suffisante à l'obtention de variantes valides, puisque parfois seules des informations plus contextuelles et liées au domaine d'expertise peuvent permettre de trancher. La priorité est ainsi donnée au rappel. Dans un deuxième temps, il s'agit d'exprimer ces critères linguistiques sous forme de contraintes assimilables par le système en proposant la définition de traits minimaux. Là encore, le problème est simplifié puisqu'on n'est pas en mesure de calculer la représentation argumentale des séquences mais seulement d'en capter quelques propriétés. Cette démarche est grandement facilitée par le dispositif de reconnaissance : les métarègles ne sont pas conçues pour générer les variantes mais pour les reconnaître dans un corpus, qui joue ainsi un rôle de filtre.

Cette expérience m'a permis de faire un premier pas en direction d'une approche plus inductive des phénomènes linguistiques, en proposant certes une modélisation *a priori*, qui applique des principes de sélection argumentale, mais qui a été identifiée sur la base d'observations préalables, et qui est conçue en combinaison avec la capacité du corpus à compenser ses tendances à la surgénération. Ce travail m'a également permis de démarrer la réflexion sur l'annotation et l'évaluation, en mettant en œuvre un calcul d'accord inter-annotateurs (mesure de *kappa*) et en m'intéressant à une tâche sémantique complexe, difficile à définir et à évaluer.

Concernant la question de la paraphrase, reformulée ici en termes de recherche de variantes terminologiques, cette expérience montre qu'une approche guidée par les corpus

et les objectifs de traitement automatique amène à considérer des segments paraphrastiques non standards au regard des définitions linguistiques. Les frontières des variantes candidates sont déterminées par la présence des mots du terme initial, sans qu'on ait nécessairement affaire à des syntagmes bien formés. Le jugement lié à la stabilité du contenu informationnel ne nous amène pas à travailler strictement sur des unités correctement délimitées sur le plan lexical et syntaxique.

Dans cette première expérience, le lien nom-verbe était déjà établi et limité aux cas de dérivation morphologique. La recherche portait sur la mise en rapport de séquences plus larges les incluant. Dans le second projet dont je rends compte, nous nous sommes focalisés cette fois sur l'identification du lien entre lexèmes nominaux et verbaux, dans le but de récolter des liens sémantiques qui ne relèvent pas de la morphologie et qui peuvent être exploités en recherche d'information.

2.3 Acquisition de lexiques sémantiques basés sur le lien nom-verbe

Je présente ici des travaux menés dans le cadre d'un projet international coordonné par Pascale Sébillot [4], dont l'objectif était de mettre au point une méthode pour acquérir automatiquement des lexiques sémantiques incorporant des liens nom-verbe, dans le but d'augmenter les capacités des systèmes de recherche d'information en étendant les possibilités de reformulation des requêtes. Ce projet a donc été défini dans le prolongement immédiat de mes résultats de thèse, mais il illustre le tournant méthodologique précédemment évoqué, c'est-à-dire le recours plus systématique aux corpus pour faire émerger les propriétés à prendre en compte dans la description des phénomènes.

2.3.1 Les hypothèses

L'objectif du projet était le développement de ressources lexicales pour le français, adaptées à la tâche de recherche d'information. Deux critères ont été retenus : tout d'abord nous avons opté pour une méthode d'acquisition à partir de corpus, pour limiter le coût de création de la ressource, pour exploiter les vertus heuristiques de cette approche inductive, et pour garantir l'adéquation entre l'information lexicale disponible et la base de textes considérée pour effectuer les recherches. C'est une alternative aux ressources externes et hors domaine utilisées systématiquement à l'époque sur l'anglais, avec le recours quasi inévitable à la base lexicale Wordnet, dont nous avons déjà évoqué certaines limites en 1.3.1. Ensuite, nous avons cherché à diversifier la nature des relations exploitables et à rompre avec une vision cloisonnée du lexique qui consiste à ne considérer dans ce type d'études que les noms, et à ne prendre en compte que des relations lexicales très étroitement balisées

4. Action de recherche partagée AUPELF-UREF (Association des Universités Partiellement ou Entièrement de Langue Française – Université des Réseaux d'Expression Française), réseau Francophone de l'Ingénierie de la Langue (Francil). Le projet réunissait des chercheurs des universités de Rennes (Pascale Sébillot et Vincent Claveau), Genève (Pierrette Bouillon), Bruxelles (Laurence Jacqmin) et Toulouse (moi-même).

(essentiellement, la synonymie et l'hyperonymie, parce qu'elles sont justement accessibles dans Wordnet). Nous avons alors décidé de nous focaliser sur l'acquisition de liens nom-verbe, de surcroît plus faciles à acquérir que des liens de nature paradigmatique, puisque nous nous sommes focalisés sur des rapports de sens qui peuvent être détectés par la recherche de cooccurrences du premier ordre. Ce lien nom-verbe n'est cette fois pas contrôlé par une relation morphologique, mais seulement par une relation sémantique, telle qu'elle est définie dans le cadre du lexique génératif. Le formalisme du lexique génératif n'ayant pas donné lieu au développement de lexiques, notre objectif était donc de construire une ressource lexicale permettant de remplir la structure *qualia* pour les noms d'un corpus, donnant ainsi accès à des composants essentiels de la sémantique nominale décrits sous la forme d'une information verbale, comme on l'a vu section 1.3.2.1.

2.3.2 La méthode

La méthode mise au point a consisté à acquérir les paires N-V d'un corpus qui sont pertinentes du point de vue sémantique, et à les distinguer d'autres paires également présentes mais qui ne comportent pas la relation sémantique visée. Nous avons considéré comme pertinentes les paires qui sont interprétables en termes de rôles *qualia*, qu'il s'agisse du rôle télique (*livre*/*lire*) ou agentif (*livre*/*écrire*). L'approche choisie dans le projet, mise en œuvre par Vincent Claveau (Claveau, 2003), est un apprentissage de type symbolique, la programmation logique inductive (PLI). Celle-ci consiste à apprendre un concept à partir d'exemples positifs et d'exemples négatifs. Plus précisément, dans les termes de la PLI, le programme est conçu pour apprendre des clauses suffisamment générales pour couvrir l'ensemble des exemples positifs, tout en couvrant le moins possible d'exemples négatifs. Cette démarche a le grand intérêt de produire des règles contextuelles, qui permettent non seulement de découvrir des instances du concept, mais d'obtenir des caractérisations de leurs contextes d'occurrence, améliorant ainsi la connaissance du phénomène en question. En d'autres termes, elle fournit une « théorie » fondée sur les données en produisant une explicitation des principes qui ont guidé l'apprentissage.

Dans une première étape du projet, les exemples positifs fournis au programme sont constitués de couples nom-verbe en contexte qui ont été validés comme instances de rôles *qualia*. Le corpus sur lequel a porté l'expérimentation est un manuel de maintenance d'hélicoptère. C'est un corpus convenable vis-à-vis de la tâche considérée dans la mesure où il contient beaucoup de noms concrets désignant des objets décrits d'après leur mode d'utilisation ou de conception : les dimensions télique (finalité, fonction) et agentive (mode de création) sont donc bien présentes. L'annotation des exemples (E+ vs E-) s'est faite de la façon suivante : nous n'avons retenu que les phrases du corpus qui contenaient au moins un nom et un verbe, puis nous avons conservé pour chaque nom les 10 verbes présentant la plus forte valeur de $Chi2$. Les couples ont été annotés manuellement hors contexte de façon binaire, selon qu'ils se conformaient ou pas à un rôle *qualia*. Par exemple, les couples (*vis*, *serrer*), (*disjoncteur*, *déclencher*) ont été considérés comme des instances de rôle *qualia,* ce qui n'est pas le cas des couples (*température*, *pouvoir*) ou (*allumage*, *réunir*). Les couples ont été ensuite reprojetés dans le texte, et les contextes calculés après un contrôle ma-

nuel qui vérifie que les couples une fois identifiés en contexte entretiennent bien la relation envisagée hors contexte.

Les informations contextuelles accessibles sont les formes et certaines propriétés morpho-syntaxiques des unités, dont les catégories grammaticales, le corpus ayant été préalablement annoté automatiquement. Nous avons donc cherché à déterminer ce qui, dans le contexte ainsi décrit, distinguait les couples pertinents des autres, ce qui revient à faire l'hypothèse (dans un premier temps) qu'une représentation très surfacique du contexte peut permettre de discriminer une relation sémantique du type de celles que décrit la structure *qualia*. Les exemples positifs fournis au système ont la forme suivante :

POSITIF(catégorie avant N, catégorie après N, catégorie avant V, type de V, distance, position)

On exclut de la description du contexte les adjectifs, les adverbes, les déterminants. Un exemple positif particulier est :

POSITIF(VRBINF,P_DE,VID,VRBINF,0,POS)

Cette clause signifie qu'un couple verbe-nom valide a les caractéristiques suivantes : le verbe est à l'infinitif (VRBINF), il précède le nom (POS). La valeur 0 pour la distance indique qu'aucun autre verbe ne s'intercale entre les deux mots. Le nom est précédé d'un verbe à l'infinitif et suivi de la préposition *de* (ex : *ouvrir la porte de*).

Le choix des éléments de contexte à considérer est très difficile à faire et ne peut en fait être justifié que par une mesure de la qualité de l'apprentissage qu'ils permettent d'obtenir, d'où de nombreuses tentatives successives, des tâtonnements. Ce travail initial a été ensuite prolongé par plusieurs expériences visant à enrichir la caractérisation du contexte en exploitant différents niveaux d'annotation, et en particulier une annotation sémantique des noms et des verbes à partir d'un typage sémantique dédié à l'étiquetage du vocabulaire du corpus (Bouillon *et al.*, 2001).

Une fois les exemples constitués, la phase d'apprentissage peut se dérouler. Il faut pour cela indiquer au programme les classes de définition de mots qu'il peut utiliser dans son calcul de généralisation : par exemple, les lemmes des prépositions peuvent être généralisés en catégories de type PREP_LIEU, PREP_DE, etc. Une partie du corpus a servi à l'apprentissage, le reste à la validation.

2.3.3 Les résultats

Nous nous intéressons ici aux résultats empiriques [5], c'est-à-dire à l'évaluation linguistique des nouveaux couples trouvés sur la partie du corpus dédiée à la validation. Au-delà de la capacité de la méthode à correctement identifier les couples pertinents, l'intérêt de l'expérience est qu'elle fournit des éléments explicatifs susceptibles d'aider à caractériser linguistiquement le phénomène visé. Je commence par expliquer les clauses produites, avant d'indiquer les principaux éléments de description que nous avons recueillis.

5. Pour une évaluation dite théorique du programme, visant par exemple à apprécier le degré de couverture des règles, cf. (Claveau, 2003).

Premier exemple de clause

POSITIF(A,B,C,D,E,F) :- VIDE(A), AUX_ETRE(C), PREPOSITIONDE(B), PRES(E).

Cette clause signifie que tout couple N-V est une instance de rôle *qualia* s'il présente les caractéristiques suivantes : le nom n'a pas de contexte gauche (VIDE(A)) et est suivi par la préposition *de* – quels que soient le type du verbe et la position respective des deux mots (puisque les paramètres D et F ne sont pas instanciés). Les deux mots ne sont pas séparés par un verbe (PRES(E)). Cette clause couvre par exemple l'E+ suivant correspondant au couple (*écrou*, *serrer*) :

(20) Les écrous des tuyauteries hydrauliques seront serrés sans exagération.

qui est décrit ainsi par la clause POSITIF(VID,P_DE,ETRE,VERBADJ,0,NEG).

Deuxième exemple de clause

POSITIF(A,C,C,D,E,F) :- PREPOSITIONLIEU(A), VIDE(C), VERBINF(D),PRES(E)

Cette clause signifie que tout couple N-V présentant les propriétés suivantes est une instance de rôle *qualia* : le verbe est à l'infinitif, le nom est précédé d'une préposition de lieu, les deux mots n'étant pas séparés par un verbe (PRES(E)). Cette clause couvre par exemple l'E+ suivant correspondant au couple (*brancher*, *prise*) :

(21) Brancher les connecteurs sur les prises électriques.

qui est décrit ainsi : POSITIF(P_SUR,VID,VID,VERBINF,0,POS).

Troisième exemple de clause

POSITIF(A,B,C,D,E) :- VIDE(A), ETAT_VERBE(C), MODALITE(B)

Cette clause, contrairement aux deux précédentes, ne s'appuie pas seulement sur les résultats de l'étiquetage morpho-syntaxique mais incorpore des étiquettes sémantiques. Elle signifie que tout couple N-V présentant les propriétés suivantes est une instance de rôle *qualia* : le nom est en début de phrase, il est suivi d'un verbe modal, et le verbe est précédé d'un verbe d'état. Cette clause couvre par exemple les E+ suivants correspondant aux couples (*platine*, *déposer*) et (*train*, *sortir*) :

(22) Les platines doivent être déposées ou le train peut être sorti à l'aide de l'électropompe de secours.

Outre les indications relatives à la proximité des deux unités et à leur position respective, les principales caractéristiques apprises sont les suivantes :

- certaines constructions syntaxiques sont favorisées : la tournure passive, la présence d'une complétive, le verbe à la forme infinitive,
- la position initiale du nom est privilégiée,
- les verbes modaux sont de bons indicateurs,
- le type sémantique de la préposition est pris en compte : la préposition indique la manière – *fixer avec leurs vis* – ou le but – *pour emmancher l'arbre d'entraînement.*

Le programme a donc bien capté la spécificité du corpus, et en particulier le format de rédaction des instructions, qui sont très riches en informations de type procédural.

L'ensemble de ces travaux a fait l'objet de différentes publications, en particulier (Bouillon *et al.*, 2000) et (Bouillon *et al.*, 2001). Ils ont été finalisés après la fin du projet par V. Claveau dans sa thèse, qui fournit le bilan le plus complet de ce travail (Claveau, 2003).

2.3.4 Bilan de cette expérience

Comme dans l'expérience précédente, le travail de description linguistique a été guidé par des catégories linguistiques qui préexistaient à la démarche d'exploration du corpus. Il s'agissait précédemment d'utiliser les propriétés de sous-catégorisation pour contrôler la sémantique des unités que l'on traitait ; ici, nous avons appliqué les propositions issues d'un modèle lexical théorique pour déterminer l'existence d'un lien sémantique entre deux mots. En ce sens, ce travail relève d'une approche résolument *corpus-based* (cf. section 9.2.2) : l'annotation des exemples s'est faite conformément à des catégories pré-établies. Néanmoins, la procédure d'annotation elle-même, en nous amenant à nous confronter à des exemples issus du corpus, et à élaborer des critères communs pour s'assurer de la cohérence de nos annotations, nous a conduit à confronter les catégories choisies à la réalité des usages, et à les préciser. Par exemple, pour déterminer l'existence d'une relation télique, nous avons mis au point des tests pour encadrer l'annotation (formulations du type « Un N c'est typiquement quelque chose que l'on V / qui sert à V »). Par ailleurs, et surtout, le processus d'apprentissage a permis de découvrir, cette fois de façon inductive, de quelle manière ces catégories s'instancient, avec des régularités qui sont celles du corpus que l'on traite, et dont on voit bien qu'elles auraient été difficiles à percevoir à l'œil nu.

Mon rôle dans un projet comme celui-ci s'est situé à plusieurs niveaux : l'expertise linguistique est requise en amont de l'apprentissage, dans la phase de définition de la tâche, qui a consisté à déterminer le type d'information qu'il s'agit d'acquérir, et conjointement dans la phase d'annotation des données qui vont servir de base à l'apprentissage. Ensuite, la mise au point du programme proprement dit passe par la définition et le réglage des éléments contextuels à prendre en compte, et, pour une des étapes de l'expérimentation, la projection d'un étiquetage sémantique adapté au corpus. En aval, l'analyse des résultats s'appuie sur l'interprétation des règles informée par la connaissance des caractéristiques du corpus traité.

Cette expérience allait dans le sens du développement de ressources lexicales aptes à répondre aux besoins spécifiques des systèmes de recherche d'informations, qui requièrent l'accès à des relations sémantiques plus variées et mieux adaptées à la base de textes explorée. V. Claveau et P. Sébillot ont par la suite prolongé et finalisé ce travail dans ses aspects applicatifs en montrant l'impact positif de ces ressources pour réaliser l'expansion de requêtes en RI (Claveau *et al.*, 2004), validant ainsi la démarche de recensement de relations sémantiques intercatégorielles.

2.4 Conclusion

J'ai insisté sur le fait que ce double travail m'a amenée à tester différentes modalités de travail sur les corpus, articulant des approches déductives et inductives. J'ai découvert à différents niveaux la dimension inductive ou tout au moins heuristique des approches sur corpus, lorsqu'elles sont outillées par des procédures élaborées de repérage et d'apprentissage. Dans les deux cas j'ai été confrontée aux difficultés de définir et d'évaluer des tâches d'annotation sémantique, et j'ai découvert certaines méthodes permettant de réduire en partie cette difficulté (mise au point de critères précis, calcul d'une mesure d'accord). Sur la question de la proximité sémantique, ces deux expériences ont confirmé l'intérêt de prendre en compte des relations intercatégorielles pour alimenter des tâches de repérage de la proximité sémantique – variation terminologique et recherche d'information, que ces relations soient ou non doublées d'une proximité morphologique, et qu'elles reposent sur des affinités de second ordre ou de premier ordre. Ces premières expériences m'ont conduite par la suite à aller plus avant dans une approche véritablement fondée sur les corpus pour faire émerger des rapports de proximité sémantique entre les mots. C'est l'objet des deux derniers chapitres de cette première partie, consacrés aux méthodes d'analyse distributionnelle automatique.

Chapitre 3

L'approche distributionnelle

La démarche harrissienne (Harris *et al.*, 1989) consiste à classer les mots de façon inductive sur la base de leurs caractéristiques distributionnelles. Plus précisément, elle considère la similarité des contextes syntaxiques dans lesquels deux mots apparaissent comme l'indice de leur proximité sur le plan sémantique (« setting up word classes in respect to their grammatical combinations » (*op.cit.*, p. 8)). Elle s'inscrit dans le cadre de l'analyse des sous-langages, dont les propriétés – et en particulier la netteté des restrictions de sélection qui régissent les relations entre opérateurs et arguments – garantissent la possibilité de mettre au jour des classes de mots bien circonscrites et caractérisables. L'hypothèse harrissienne et la prospérité qu'elle a connue dans le cadre des recherches en TAL, est présentée en particulier dans Habert et Zweigenbaum (2003). Son automatisation a donné lieu à ce que Habert et Nazarenko (1996) appellent des « versions affaiblies » de la méthode – les rapprochements entre mots sont effectués sur la base de simples critères de cooccurrence (depuis Church et Hanks (1990)) – ou à des implémentations plus conformes au modèle harrissien, basées sur le calcul des contextes syntaxiques partagés. C'est à ce deuxième cas de figure que je me suis intéressée, à travers mes participations à deux projets de mise en œuvre de l'analyse distributionnelle automatique réalisés pour le traitement du français, ZELLIG puis UPERY.

Les deux expériences que je relate ont été menées à quelques années d'intervalle. Le travail avec B. Habert autour de ZELLIG a été réalisé en 1997, les sorties du programme UPERY sur lesquelles j'ai travaillé ont été construites par D. Bourigault à partir de 2002. Les hypothèses de recherche sont très proches dans les deux cas. Par contre, en quelques années, leur mise en œuvre a changé radicalement du fait du développement d'un analyseur syntaxique à large couverture, SYNTEX, et de la disponibilité de corpus de grande taille pour le français.

Mon cheminement sur cette période peut être résumé ainsi : avec B. Habert, je me suis intéressée à la possibilité de déplacer le champ d'expérimentation de l'ADA, et nous avons confronté ZELLIG à l'analyse de corpus non techniques. Les premiers résultats nous ont convaincus à la fois de l'intérêt potentiel de l'approche, et de l'inadéquation des outils dont nous disposions en amont, aussi bien que des ressources textuelles que nous traitions. La collaboration avec D. Bourigault sur SYNTEX puis UPERY m'a permis de participer au

développement d'un système plus performant. Dans une dernière étape, qui se poursuit jusqu'à aujourd'hui, je me suis intéressée à l'utilisation et à la valorisation des résultats d'UPERY : mon objectif est de tirer parti de la ressource produite - les voisins distributionnels - et de comprendre quel est l'apport de l'analyse distributionnelle automatique, pour le TAL et pour la linguistique, comme mode d'acquisition d'informations sémantiques à partir de grands corpus de textes non spécialisés.

Sur le plan de ma contribution personnelle, elle est similaire dans les deux expériences. J'ai été associée à la réflexion sur les outils d'analyse et à leur développement (Habert et Fabre, 1999; Bourigault et Fabre, 2000), mais j'interviens pour l'essentiel en aval des traitements réalisés : je me suis consacrée à l'analyse, l'évaluation et la validation des rapprochements calculés par les programmes. La conviction qui m'a animée est que ces données, élaborées d'abord pour assister la constitution de réseaux terminologiques (Habert et Fabre, 1999; Bourigault et Lame, 2002) pouvaient intéresser plus largement les travaux de linguistique et de TAL sur le lexique général.

Dans ce chapitre, je commence par présenter à grands traits les deux systèmes d'ADA que j'ai utilisés, en les replaçant dans le contexte des expériences similaires qui sont menées au niveau international sur d'autres langues (3.1). Je présente ensuite les travaux que j'ai menés autour de ZELLIG, et je montre comment les résultats d'UPERY ont fait évoluer cette réflexion initiale. En 3.3, je présente les pistes de travail qui justifient selon moi une démarche d'analyse et de valorisation de cette ressource, et je mets l'accent sur les difficultés méthodologiques rencontrées. En 3.4, je reviens sur la question des relations intercatégorielles, que cette ressource permet d'alimenter avec des données nouvelles. Enfin, je présente en 3.5 une première tentative d'utiliser la ressource dans le contexte applicatif de la recherche d'information.

3.1 Deux expériences d'automatisation de l'hypothèse distributionnelle

ZELLIG ZELLIG est un programme conçu par B. Habert (Habert et Nazarenko, 1996), qui permet de construire et de visualiser des regroupements de mots calculés à partir des contextes syntaxiques que ceux-ci partagent. Ces liens syntaxiques sont calculés en amont par l'analyseur LEXTER, prédecesseur de SYNTEX, également développé par D. Bourigault. L'approche est alors totalement nouvelle sur le français. Elle s'inspire de travaux antérieurs menés par (Grefenstette, 1994b), visant l'acquisition automatique de relations entre mots pour assister le développement de thésaurus. Les regroupements effectués par ZELLIG sont de deux types :

- composantes connexes : elles regroupent des mots qui partagent des contextes communs 2 à 2 (il y a toujours un chemin dans le graphe entre 2 nœuds donnés) ;
- cliques : elles regroupent des mots qui partagent des contextes communs avec tous les autres mots de la clique (chaque nœud est relié avec les autres par un arc).

Les regroupements sont calculés sur la base de dépendances syntaxiques élémentaires, qui sont des relations binaires entre mots pleins lemmatisés. Ces dépendances binaires sont

calculées à partir de groupes nominaux analysés automatiquement (réalisant la phase de normalisation également spécifiée par Harris). Par exemple, la phrase suivante (l'exemple est repris de Habert et Fabre (1998) et tiré du corpus MITTERRAND, décrit section 3.2.1) :

(1) *J'ai toujours dit également que la connaissance de l'OLP comme organisation représentative du peuple palestinien était soumise à certaines conditions.*

fournit deux groupes nominaux analysés par LEXTER : *la reconnaissance de l'OLP* et *organisation représentative du peuple palestinien* qui se décomposent eux-mêmes en une série de dépendances binaires : (*reconnaissance, OLP*), (*organisation, représentative*), (*organisation, peuple*), (*peuple, palestinien*).

Ces dépendances binaires constituent les distributions des mots qui permettent d'effectuer des regroupements. Par rapport à des traitements basés sur le calcul de simples cooccurrences de mots dans une fenêtre de taille donnée, ces regroupements s'appuient sur des associations validées syntaxiquement. ZELLIG construit ensuite un graphe dont les nœuds sont les lemmes et dont chaque arc figure l'ensemble des contextes qui relient deux lemmes. Ces dépendances sont limitées au domaine nominal dans la mesure où LEXTER effectue des analyses partielles, ciblées sur les seuls groupes nominaux, en raison de leur fort potentiel terminologique.

UPERY La démarche mise en œuvre par le programme UPERY (Bourigault, 2002) est sensiblement la même. Cette fois, l'analyseur en amont est SYNTEX, ce qui permet d'augmenter considérablement la qualité et le volume des contextes exploitables puisque l'analyseur traite les principales relations syntaxiques au sein de la phrase. On retrouve les mêmes principes de réduction et de normalisation syntaxique, réalisés de façon plus conséquente (transformation du passif, réduction des coordinations, résolution de l'antécédence relative, etc.). Les unités rapprochées peuvent être des unités complexes. Les relations de dépendance sont traduites en triplets prédicat-relation-argument.

Le programme d'analyse distributionnelle présente cependant une particularité qui le distingue d'autres systèmes semblables. Il réalise en effet deux types de rapprochement : les arguments sont rapprochés sur la base des couples lemme-relation qu'ils partagent, alors que les prédicats, considérés comme une séquence lemme-relation, sont rapprochés sur la base des contextes arguments qu'ils ont en commun. UPERY rapproche donc des voisins-prédicats, et des voisins-arguments. Le tableau 3.1 illustre ces deux vues sur un même mot à partir de l'exemple du mot *réforme*. Le corpus dont sont issues ces données est le corpus LM10 décrit plus loin (section 3.2.1). En tant qu'argument, *réforme* se rapproche d'autres noms sémantiquement proches comme *projet*, *changement*, *mesure*, sur la base de contextes partagés du type *voter*_CONTRE ou *mettre en chantier*_OBJ. Le même nom *réforme*, cette fois considéré en position de prédicat, apparaît principalement (68% des cooccurrences) avec la relation DE [1]. Le prédicat *réforme*_DE a pour voisins d'autres prédicats, qui peuvent comporter soit des noms (*modification*_DE, *refonte*_DE) soit des verbes (*réformer*_OBJ,

1. Les autres relations qui permettent d'identifier des voisins sont MOD (qualification adjectivale : *réforme souhaitable*), et marginalement EN (*réforme en discussion*) et À (*réforme au sénat*).

*modifier*_OBJ). Les relations de voisinage peuvent donc varier de manière importante selon le type de rapprochement effectué.

réforme en tant que prédicat	*réforme* en tant qu'argument
*réformer*_OBJ	*projet*
*modification*_DE	*changement*
*refonte*_DE	*mesure*
*fonctionnement*_DE	*réduction*
*modifier*_OBJ	*programme*

TABLE 3.1 – Vue sur un même mot considéré comme prédicat ou argument

Ce mode de rapprochement n'est certainement pas optimal si l'on fait l'hypothèse qu'un classement est d'autant meilleur qu'il est obtenu en prenant en compte le plus grand volume de données possible. En effet, en distinguant différents types de rapprochement entre mots (entre arguments et prédicats, et, au sein des prédicats, entre différentes relations), on éparpille les cooccurrents, ce qui peut avoir deux effets néfastes : certains rapprochements ne sont pas apparents parce qu'ils ne dépassent pas les seuils imposés ; l'importance de certains voisins peut être sous-évaluée, car on ne fait pas la différence entre les voisins qui émergent simultanément dans différentes configurations et ceux qui n'apparaissent que pour une relation syntaxique donnée. On peut remédier au deuxième problème en aval du traitement effectué. L'intérêt ici est que cette approche différenciée des contextes ouvre des possibilités d'observation linguistique inédites, comme nous allons le voir.

Plusieurs mesures de similarité ont été implémentées par D. Bourigault (Bourigault, 2002). L'indice de Jaccard a d'abord été employé. A étant l'ensemble des contextes associés au mot M1 (prédicat ou argument), B l'ensemble des contextes associés au mot M2, cet indice estime la similarité entre A et B en calculant le rapport entre la cardinalité de l'intersection de A et de B (les contextes partagés par les deux mots) et celle de leur union.

Plus précisément :

$$jaccard = \frac{a}{n1+n2-a}$$

où a est le nombre de contextes partagés, *n1* le nombre de contextes de M1, *n2* le nombre de contextes de M2. La version du programme que j'utilise intègre également la mesure de (Lin, 1998b), qui ajoute une mesure de probabilité sur les contextes, de façon à prendre en compte leur degré d'informativité : les contextes plus rares sont considérés comme de meilleurs indices de similarité.

Travaux associés De nombreux travaux ont recours à l'hypothèse de distributionalité pour acquérir automatiquement des informations sémantiques à partir de corpus, depuis (Grefenstette, 1994b) et (Lin, 1998a). Cette démarche a connu un nouveau regain pendant la période récente. Baroni et Lenci (2009) ont proposé un état des lieux très utile. Ils distinguent deux phases de l'analyse distributionnelle : la première consiste à construire une « mémoire sémantique distributionnelle » (*distributional semantic memory*) que l'on

peut représenter par un graphe reliant les mots entre eux via des relations syntaxiques. Les mots ainsi que les relations considérés peuvent être de nature variable (comme illustré à travers les expériences menées avec ZELLIG et UPERY). Les arcs sont pondérés par un indice de type information mutuelle. Ce graphe fournit une liste de triplets mot-relation-mot, ou, selon la terminologie de Baroni et Lenci, *concept-link-concept*. Cette mémoire peut ensuite être exploitée de diverses manières, par le biais de « vues sémantiques » (*semantic views*) spécifiques, dont Baroni et Lenci montrent qu'elles peuvent être mises en rapport avec des tâches sémantiques particulières. Des trois vues présentées par les auteurs, deux sont expérimentées dans UPERY, une dans ZELLIG [2].

- *concept+concept-by-link* (CCxL) : c'est l'option suivie par la quasi-totalité des systèmes. Elle consiste à classer les mots sur la base des relations qu'ils partagent. Elle correspond aux rapprochements effectués par ZELLIG, et par UPERY pour les arguments. Le tableau 3.2 montre [3] un petit échantillon des contextes similaires dans lesquels apparaissent les noms *avis* et *recommandation* (objets de *rédiger*, compléments en *à* de *se conformer*, expansions prépositionnelles du nom *publication*), assortis de la valeur d'information mutuelle (calculée dans le corpus LM10, décrit section 3.2.1).

	rédiger_OBJ	se conformer_À	publication_DE
avis	3.289	4.974	3.492
recommandation	4.132	6. 474	2.958

TABLE 3.2 – Rapprochement d'arguments avec UPERY

- *concept+link-by-concept* (CLxC) : cette configuration rapproche les positions argumentales des mots prédicatifs. C'est l'option choisie dans UPERY pour effectuer le rapprochement entre prédicats. Le tableau 3.3 fait état d'une proximité entre les objets du verbe *rédiger* et ceux du nom *publication*, apparaissant via son expansion prépositionnelle en *de*.

	avis	*recommandation*	*biographie*
*rédiger*_OBJ	3.289	4.132	5.492
*publication*_DE	3.492	2.958	4.442

TABLE 3.3 – Rapprochement de prédicats avec UPERY

2. La troisième possibilité (*link-by-concept+concept* (LxCC)) correspond à une recherche de relations analogues, rapprochées parce qu'elles unissent les mêmes paires de concepts.

3. Sauf indication contraire, les exemples sont tirés du traitement par UPERY du corpus LM10 (3.2.1).

Cette différence dans le traitement des prédicats et des arguments a été, au départ, dictée par l'hypothèse d'un fonctionnement différent d'un même mot (nom ou adjectif), sur le plan sémantique, selon qu'il est prédicat ou argument. Mais elle présente un autre intérêt manifeste, celui de diversifier les modes de rapprochement entre mots (cf. section 3.3.2). Une alternative consisterait à traiter tous les mots selon le modèle CCxL, en utilisant le principe des relations inverses (notées r-1) (Baroni et Lenci, 2009). C'est ce que font classiquement les programmes d'analyse distributionnelle : « For each word a context vector is constructed that consists of all grammatical relations a word is found in with the accompanying word attached to it. » (Bouma *et al.*, 2005).

Dans le cas des trois prédicats considérés dans le tableau 3.3, et en limitant l'illustration au seul argument *avis*, on constate qu'ils n'ont cette fois aucun contexte commun, puisque la méthode ne permet de rapprocher que des prédicats qui partagent les mêmes relations, ce qui limite les possibilités de regroupements. En maintenant la relation du côté de la variable et non du contexte, on rend en particulier possibles les rapprochements intercatégoriels. Le travail sur les couples de voisins V-N (section 3.4) vise à tirer parti de cette singularité du programme UPERY [4].

	OBJ-1_*avis*	À-1_*avis*	DE-1_*avis*
rédiger	3.289	0	0
se conformer	0	4.974	0
publication	0	0	3.492

TABLE 3.4 – Rapprochement par le biais de relations inverses

3.2 Analyse distributionnelle de corpus non spécialisés

3.2.1 Elargir le champ d'application de la méthode

Avec ZELLIG comme avec UPERY, mon objectif a été de travailler sur l'application de la méthode à des corpus non techniques. Nous avons choisi de parler, dans les publications concernées par ces travaux, de corpus « proches de la langue générale », mais cette formulation ne me semble plus adéquate. Tous les corpus utilisés (le premier avec ZELLIG, les trois autres avec UPERY) relèvent à l'évidence de types de textes particuliers, auxquels on peut attribuer une étiquette – grossière – de genre :

Discours politiques : corpus *Mitterrand* (désormais MITTERRAND)

Ce corpus, rassemblé par Dominique Labbé, comprend l'ensemble des discours radio-

4. Une fois les prédicats rapprochés, nous avons parfois fait le choix (comme on le verra en particulier en 4.2.3) d'extrapoler l'information calculée, et de considérer que la proximité établie entre les couples lemme_RELATION s'appliquait au lemme dans son entier. Cette option demande sans doute à être mieux contrôlée.

télévisés de François Mitterrand pendant son premier septennat. Il comporte environ 300 000 mots (Fabre *et al.*, 1997).

Textes encyclopédiques : *Wikipedia* (WIKIPEDIA)
Ce corpus a été recueilli par Franck Sajous. Il contient environ 470 000 articles de la version francophone de l'encyclopédie en ligne Wikipédia, soit 194 millions de mots.

Textes journalistiques : *Le Monde* (LM10)
Ce corpus comprend 10 ans du journal *Le Monde*[5], paru entre 1991 et 2000, soit 200 millions de mots.

Textes littéraires : Frantext 20^e, (FRANTEXT20)
Ce corpus comprend la partie romans du 20^esiècle de la base textuelle *Frantext*[6].

Mais ce que l'on peut dire dans tous ces cas, c'est que ces corpus sont constitués de discours appelés à bénéficier d'une large réception et ne reposent pas sur le partage de connaissances spécialisées. Ces caractéristiques externes les distinguent donc clairement des types de textes visés par l'analyse harrissienne, et de fait ils ne présentent pas les propriétés de redondance qui facilitent – voire légitiment, c'est la question qui se pose toujours en arrière-plan de nos travaux – l'analyse distributionnelle.

3.2.2 Premières explorations avec ZELLIG

Le travail d'analyse que j'ai réalisé sur les sorties de ZELLIG a consisté à travailler sur le corpus MITTERRAND de manière à caractériser l'intégralité des regroupements obtenus par le système et à comprendre les effets du déplacement opéré par le fait de quitter le champ des corpus spécialisés pour aborder celui du discours politique. Ce travail est décrit dans (Habert et Fabre, 1998) et (Fabre et Habert, 1998).

Analyses Les expériences menées par B. Habert et ses collègues (Bouaud *et al.*, 2000) sur un corpus médical, MENELAS, avaient montré que l'analyse distributionnelle fournissait une représentation utile de l'organisation conceptuelle du domaine, à condition d'appliquer aux résultats une phase d'interprétation pour distinguer des sous-classes au sein de classes hétérogènes, pour fusionner certains groupes et pour caractériser les classes mises au jour. On retrouve les relations classiques d'hyperonymie, d'antonymie et de synonymie, et on peut étiqueter la majorité des regroupements (*interventions*, *examens*, *organes*, *affections*, ...). Dans le cas du corpus MITTERRAND au contraire, les regroupements ne sont pas caractérisables en termes de classes conceptuelles. Certaines composantes connexes sont très vastes (plus de 100 mots), ce qui atteste du peu de spécialisation des contextes distributionnels. On dégage quelques ensembles : le vocabulaire géo-politique des pays et des institutions (*gouvernement*, *ministre*, *peuple*, *région*...), des adjectifs identifiant différents niveaux de réalité (*économique*, *social*, *industriel*, *européen*...), des noms généraux caractérisant les modes d'application du politique (*plan*, *décision*, *situation*, *moyen*...). Les

5. Le corpus LM10 a été construit à partir de ressources obtenues auprès de l'agence ELDA et à l'aide de programmes de nettoyage et de balisage réalisés par B. Habert.

6. Ce corpus a été utilisé dans le cadre d'une convention passée entre l'ATILF et CLLE-ERSS.

composantes plus petites ou les cliques permettent d'identifier certains groupes homogènes, comme le vocabulaire de la force, du temps, de l'entreprise, des acteurs politiques, etc. Les cliques donnent également accès à des relations d'antonymie (*privé-public*, *riche-pauvre*), de synonymie (*français-national*, *pays-nation*), de méronymie (*pays-monde*, *ministre-gouvernement*), mais ces relations sont presque toujours partielles : par exemple, dans la clique (*mois*, *année*, *guerre*), le mot *guerre* semble un intrus, mais sa présence s'explique si on la considére à la lumière des contextes qui unissent les mots de la clique (par exemple : *cours de* ∼, *prochain*∼, ∼ *dernier*), c'est-à-dire comme un jalon temporel dans le discours présidentiel. C'est donc seulement un aspect de la sémantique du mot qui est éclairé. De façon similaire, les grandes composantes manifestent la polysémie des mots dans ce type de corpus : certaines constituent des faisceaux autour d'un mot et exhibent différentes facettes sémantiques que l'analyse distributionnelle n'est pas en mesure de distinguer. C'est le cas de mots comme *politique* et *pays*. L'examen manuel de ces faisceaux permet seul de dégager une sorte de cartographie de l'emploi de ces mots – on distingue par exemple quatre dimensions d'emploi du mot *pays* : l'humain, l'économique, le politique, le militaire. L'analyse conceptuelle, la démarche de catégorisation qui pouvait s'appliquer à MENELAS, n'est donc pas possible dans le cas de MITTERRAND. Le dispositif permet certes de dégager des classes et des relations, mais elles sont somme toute assez banales, et très mouvantes. L'information la plus riche réside dans les contextes d'emploi, dont les regroupements offrent une vision synthétique qui permet de faciliter l'étude du vocabulaire dans le discours de Mitterrand.

Conclusions Cette expérience a clairement montré que les données disponibles étaient très insuffisantes pour faire émerger des rapprochements sémantiques permettant de soutenir une analyse approfondie du discours contenu dans les textes traités par ZELLIG. Le traitement de textes de nature non technique, présentant moins de redondance, exige un volume de données plus important. Les seuils imposés (pour filtrer les erreurs de traitement en amont, et se baser sur des associations suffisamment régulières) sélectionnent en effet des contextes très peu spécifiques. Par ailleurs, la limitation du traitement aux seuls GN n'avait plus de sens : si elle pouvait être justifiée par des traitements orientés terminologie, dans le cas d'un corpus comme celui-ci la non prise en compte du système verbal réduisait considérablement la portée de l'étude.

Par ailleurs, les contextes partagés par les mots contribuent de façon très variable à leur rapprochement, ce qui nous avait amené à conclure à la nécessité d'étudier de plus près l'impact du type de contextes sur la procédure de regroupement. L'étude a montré que la présence de contextes diversifiés du point de vue syntaxique et sémantique semblait un meilleur indice de proximité sémantique. Au contraire, les relations plus ténues reposent sur des contextes plus spécialisés, ou à l'inverse trop banals. Par exemple, les mots *pays* et *nation* sont substituables dans une grande diversité de contextes dans ce discours. Par contre, les mots *décision* et *question* ont seulement en commun des contextes indiquant le recensement (*dernier* ∼, *deuxième* ∼, *nombre de* ∼, etc.), ou désignant des individus (∼ *de monsieur*, ∼ *de ministre*). L'examen des contextes montre que ces deux mots ne sont pas synonymes mais sont à placer ensemble dans des catégories spécifiques : événements

ponctuels ordonnés dans le temps et actes d'énonciation.

Ces trois points – volume des données, analyse syntaxique étendue, attention portée aux contextes – ont été pris en compte ultérieurement dans le développement du programme UPERY.

3.2.3 De nouvelles conditions d'expérience avec UPERY

Avec UPERY, les conditions d'analyse sont nettement plus favorables. Le volume de données distributionnelles est beaucoup plus important, à la fois parce que les corpus traités sont près de 1000 fois plus vastes, et parce que la diversité des relations analysées par SYNTEX est également beaucoup plus importante. Les types de relations récupérés sont les relations objet et sujet des verbes, les relations adjectivales et prépositionnelles (compléments en *de*, *à*, *sur*, *avec*, *dans*, ...). Pour le corpus LM10, le nombre de couples prédicat/argument de fréquence supérieure ou égale à 5 est de près de 3 millions, et le nombre de couples de voisins prédicats et de voisins arguments calculés à partir de ces couples est de plus de 6 millions, pour un coefficient de Lin supérieur ou égal à 0.1 [7].

Malgré cet accroissement considérable du nombre de contextes, la caractérisation des relations calculées s'avère également difficile. D. Bourigault et E. Galy ont ainsi montré que les voisins présentent un recouvrement minime par rapport à un dictionnaire de synonymes (Galy et Bourigault, 2005). L'expérience qu'ils ont menée a consisté à confronter les voisins issus des corpus LM10 et FRANTEXT20 aux synonymes recensés dans le dictionnaire électronique du CRISCO, DICOSYN . En considérant l'ensemble des voisins dont la valeur de jaccard dépasse un certain seuil (0.1), ils ont calculé (tableau 3.5) que, selon le corpus considéré, 10 à 22% des synonymes sont retrouvés dans les voisins. Mais ces synonymes constituent une part négligeable des voisins calculés (1 à 3%).

	précision	rappel
corpus LM10	1%	22%
corpus FRANTEXT20	3%	10%

TABLE 3.5 – Comparaison des voisins avec la ressource DICOSYN (sans filtrage) (Galy et Bourigault, 2005)

Ces résultats doivent être nuancés. L'étude devrait être répliquée de manière à faire varier le seuil de similarité. L'examen manuel répété des voisins montre en effet que la densité des synonymes est bien plus importante au sein des voisins les plus fortement reliés. Il serait utile de le vérifier à grande échelle. D'autres travaux, en particulier ceux menés par (Lin *et al.*, 2003) font état de chiffres nettement plus favorables, certes dans des conditions

7. Dans tous les résultats que je présente, c'est toujours la valeur utilisée par défaut, sauf indication contraire.

d'expérimentation différentes : corpus plus gros, contextes pris en compte différents, filtrage complémentaire par patrons. Néanmoins, quelles que soient les nuances que l'on peut apporter à ce résultat préliminaire, il confirme l'impression que l'on a en découvrant les voisins, à savoir que les relations sémantiques qui sous-tendent le voisinage distributionnel sont variées, et ne se limitent pas, loin s'en faut, aux relations sémantiques entre mots que l'on a l'habitude de considérer et qui forment l'ossature des bases lexicales. Mon objectif a consisté dès lors à tenter de mieux caractériser cette ressource, que j'appellerai désormais les voisins distributionnels, c'est-à-dire d'en comprendre et d'en contrôler la diversité, afin de la rendre utilisable.

3.3 Les voisins distributionnels, un observatoire des relations sémantiques

La construction de ce type de données ouvre des perspectives nouvelles qui intéressent aussi bien le TAL que la linguistique. On trouve dans cette ressource, à côté des relations lexicales traditionnelles, les différents types de relations non classiques entre mots, que nous avions présentées section 1.2. En voici quelques exemples (je ne considère ici que des rapprochements associés à un indice de similarité élevé, en l'occurrence $\geq 0{,}2$, et je fais abstraction des relations concernées lorsqu'ils s'agit d'un rapprochement de prédicats) :

Relations lexicales classiques :

synonymie : *étonnant* est voisin de *étrange*, *surprenant*, *singulier*

antonymie : *identique* est voisin de *différent*, *contradictoire*

hyperonymie : *musée* est voisin de *bâtiment*

co-hyponymie : *adolescent* est voisin de *jeune fille*, *garçon*, *gamin*

Relations non classiques :

relation thématique : *auteur* est voisin des verbes *écrire*, *lire*, *publier*

regroupement dans un même champ sémantique : *producteur* est voisin de *industrie, importation, marché, acheter* (champ des échanges commerciaux)

relations associatives : *marché* est voisin de *parquet, tribunal* (lieu de l'activité)

Mais la richesse de cette ressource en fait également la limite : ces relations se côtoient sans qu'il soit possible de les démêler en utilisant les paramètres disponibles (type de contexte, valeur de similarité). En ce sens, différentes pistes de travail se dessinent à partir des résultats du programme UPERY, pour appréhender le contenu de cette ressource, déterminer son utilité et ses contextes d'utilisation. Je fais état dans ce qui suit de quelques apports possibles pour l'analyse linguistique, et je montre de quelle manière j'ai commencé à aborder certains de ces pistes.

3.3.1 Cooccurrents syntaxiques

La base des voisins distributionnels fournit en premier lieu les cooccurrents d'un lemme considéré comme argument ou intégré à un prédicat. Les constructions qui lui sont associées sont accessibles avec des éléments de quantification relatifs à la productivité de chaque relation. L'analyse syntaxique réalisée en amont fournit en position d'argument des syntagmes aussi bien que des mots simples. Le tableau 3.6 montre par exemple la répartition des compléments prépositionnels du verbe *travailler* dans le corpus LM10. La productivité est mesurée en nombre de dépendants différents pour une relation donnée.

travailler + prép		
relation	productivité	exemples de dépendants
à	301	*renouveau, biographie, usine, Brésil, rythme*
dans	253	*mine de charbon, fabrique, cabinet ministériel, quartier difficile*
sur	245	*chantier de construction, nouveau projet, campus, théorie, toile*
pour	172	*organe de presse, CIA, propre compte, démocratie, conseil général*

TABLE 3.6 – Cooccurrents du verbe *travailler* dans LM10

On constate à la lecture de cet exemple que certaines classes de cooccurrents sont très hétérogènes puisque la relation associée à une même préposition peut être ambiguë. C'est le cas en particulier de la relation en *à* qui recouvre à la fois pour le verbe *travailler* des compléments exprimant l'objet (*travailler au renouveau*, *à la biographie*), la localisation (*travailler au Brésil*), la manière (*travailler au rythme*). C'est également le

cas, dans une moindre mesure de la relation *sur* (*travailler sur un chantier* – localisation – vs *travailler sur une théorie* – objet). Ce type d'ambiguïté, massif dans un corpus non spécialisé, est aggravé par l'incapacité de l'analyseur à discriminer entre des types de rattachement différents (complément oblique ou circonstant, cf. 5.2). Néanmoins, ces données fournissent des informations intéressantes pour assister la constitution de ressources syntactico-sémantiques de type FrameNet[8] : elles peuvent permettre de dégager les principaux éléments de construction qui se réalisent dans le corpus pour un prédicat donné (dans le cas du verbe *travailler* : objet, localisation, employeur). Mais, inévitablement, un tri manuel et scrupuleux des contextes est requis, ainsi que la prise en compte du caractère éventuellement polysémique des prédicats mis au jour. Dans le cas de *travailler*, il faut par exemple distinguer le sens « exercer une activité rémunérée » – sens A.1.b du *TLF* – que l'on trouve dans *travailler à l'usine* et le sens « faire des investigations » – sens A.1.d. – que l'on trouve dans *travailler sur une théorie*. L'approche distributionnelle appliquée à un corpus non spécialisé effectue des regroupements en escamotant certaines différences de fonctionnement syntaxique et sémantique. Il faut savoir en tenir compte lorsque l'on s'appuie sur ces données, ce qui suppose d'examiner la diversité des dépendants qui ont contribué à la mise au jour d'un prédicat, et de pouvoir retourner aux contextes d'origine.

Je reviens dans la deuxième partie de ce mémoire sur ce niveau de la cooccurrence syntaxique.

3.3.2 Rapprochement de structures prédicatives

Si l'on s'intéresse maintenant aux rapprochements de nature paradigmatique, les voisins distributionnels mettent en correspondance des structures prédicatives qui partagent les mêmes dépendants. Rappelons qu'il s'agit d'une spécificité de l'approche UPERY. Si l'on reprend l'exemple de *travailler*, on observe les rapprochements suivants :

- *travailler*_SUR est proche des prédicats *étudier*_OBJ et *concevoir*_OBJ
- *travailler*_À est proche de *s'installer*_À et *vivre*_À
- *travailler*_POUR est proche de *employer*_SUJ et *contrat*_AVEC

Le voisinage sémantique calculé au niveau du prédicat permet d'observer des distinctions de sens qui peuvent alimenter le recensement suggéré précédemment, à savoir les constructions d'un prédicat et les actants sémantiques qui lui sont associés. Ainsi, le voisinage concernant le prédicat *travailler*_POUR se comprend de la façon suivante : les dépendants nominaux qui apparaissent dans cette position recoupent fortement ceux que l'on trouve en position de sujet du verbe *employer*, ou comme expansion prépositionnelle du nom *contrat*. Les cooccurrents partagés sont en effet : *multinationale*, *firme*, *municipalité*, *elf*, etc.

Ce point d'entrée par les prédicats offre ainsi des pistes d'analyse très intéressantes, que je ne fais ici que suggérer. A titre d'exemple, on peut s'intéresser au cas particulier

8. Nous avons proposé, avec D. Bourigault et E. Galy une communication sur ce thème : « Construire des ressources sémantiques générales à partir de corpus » dans le cadre de la journée qui s'est déroulée à Nancy en 2006 : « Des ressources sémantiques existantes à un FrameNet français ? (Contre-)arguments, ressources, méthodes et outils » organisée par Suzanne Alt (ATILF, Nancy) et Guillaume Pitel (LORIA, Nancy).

du rapprochement de prédicats impliquant un même verbe (tableau 3.7, données tirées de WIKIPEDIA). Ce type de voisinage permet de relever deux phénomènes en particulier :

- des cas de synonymie entre prépositions, qui peut être limitée à un prédicat verbal particulier (*jouer sur* ou *de*) ou présenter une certaine généralité et rassembler plusieurs couples de prédicats (*comme* / *en tant que*).
- des cas d'alternance syntaxique : un même rôle sémantique est porté par des positions sémantiques différentes (*la productivité augmente* vs *on augmente la productivité*, *on habite un appartement* vs *dans un appartement*).

Voisins	Exemples de cooccurrents
*augmenter*_OBJ / *augmenter*_SUJ	*productivité*, *fréquentation*, *niveau de vie*
*habiter*_OBJ / *habiter*_DANS	*appartement*, *environs*, *zone*
*jouer*_SUR / *jouer*_DE	*la confusion*, *l'ambiguïté*, *l'humour*
*succéder*_COMME / *succéder*_EN TANT QUE	*chef de parti*, *directeur*, *professeur*

TABLE 3.7 – Rapprochements de prédicats pour un même verbe

On peut également étudier la mise en correspondance de structures syntaxiques non symétriques entre différents verbes, comme illustré dans le tableau 3.8, qui montre par exemple une proximité entre des prédicats intégrant une relation objet, et des prédicats intégrant une préposition (exemple : *le contrat* (*porte sur* / *concerne*) *les fournitures.*)

Voisins	Exemples de cooccurrents
*concerner*_OBJ / *porter*_SUR	*fourniture*, *sexualité*, *modalité*
*analyser*_OBJ / *réfléchir*_À	*contenu*, *impact*, *stratégie*
*intervenir*_AUPRÈS DE / *prendre contact*_AVEC	*autorité*, *chef*, *commission*
*se consacrer*_À / *se concentrer*_SUR	*tâche*, *métier*, *préparation*

TABLE 3.8 – Rapprochements de prédicats par des relations différentes

Ce type de données fournit des éléments intéressants pour le recensement des liens de synonymie. Les dictionnaires de synonymes ont tendance à rapprocher les verbes sur la base d'une équivalence de construction syntaxique [9], peut-être en lien avec l'hypothèse d'une correspondance entre classification sémantique et « schèmes syntaxiques » (Dubois et Dubois-Charlier, 1997), ou tout simplement pour simplifier la description en n'introduisant

9. Ce n'est à l'évidence pas toujours vrai, et il faudrait vérifier cette hypothèse. Par exemple, un des dictionnaires compilés par DICOSYN recense le couple *porter sur* / *concerner*.

pas de référence à la complémentation. A titre d'exemple, le verbe *analyser* a principalement des synonymes transitifs dans DICOSYN (à l'exception de *rendre compte*). L'examen des voisins permet de repérer d'autres correspondances, comme *s'interroger_*SUR, *se pencher_*SUR, *réfléchir_*À, *revenir_*SUR (jaccard>=0,3). Par ailleurs, ce type de données fournit également des clés pour la désambiguïsation. Le verbe *porter* n'est synonyme de *concerner* que dans le sens correspondant à l'emploi *porter_ sur*. D'autres verbes seront au contraire voisins à travers une large palette de relations.

Avec ces données, nous sommes au coeur de la question de l'équivalence sémantique dont nous avons discuté dans le chapitre précédent. Les relations de proximité calculées pour les couples relevant des deux cas de figure illustrés dans les tableaux 3.7 et 3.8 sont précisément de l'ordre de ce que (Ibrahim *et al.*, 2003) cherchaient à capter à partir de l'exploitation de corpus alignés (section 1.1.2). Ce type de données semble pouvoir constituer un très bon point de départ pour l'acquisition de paraphrases structurelles, en faisant l'économie du recours à des corpus alignés ou comparables. Il y a là une piste importante à explorer.

3.3.3 Accès à des liens sémantiques diversifiés – un exemple

On peut s'étonner que le recouvrement entre les voisins distributionnels et un dictionnaire de synonymes soit si faible. Pour en prendre la mesure, je propose de dérouler un exemple qui permettra d'illustrer la comparaison entre les deux sources d'informations. Cet examen très sommaire vise seulement à initier la réflexion sur l'apport potentiel des voisins pour la description lexicographique, et la difficulté à évaluer une telle ressource.

Cet exemple concerne le verbe *protester*, et compare les relations recensées dans la base des voisins (tirés de LM10) et dans le dictionnaire DICOSYN, en considérant dans celui-ci à la fois les synonymes et les antonymes. Je propose de ne considérer ici que le prédicat *protester_*CONTRE, qui regroupe à lui seul 77 % des cooccurrents du verbe, comme on le voit sur le tableau 3.9. Ce tableau permet de montrer également que certaines relations peuvent présenter peu de cooccurrents, et néanmoins générer beaucoup de voisinage. C'est le cas, pour cet exemple, des relations AUPRÈS DE et SUJ. Les cooccurrents correspondent à des référents humains – individus ou collectifs – (*ambassadeur*, *ministre*, *président*, *gouvernement*) très fréquents dans le corpus. Les contextes partagés par un grand nombre de mots, peu discriminants, ont tendance à amorcer des relations de proximité moins pertinentes.

Le prédicat *protester_*CONTRE a 389 voisins. Le verbe *protester* possède 68 synonymes ou antonymes dans DICOSYN. Si l'on considère le rappel, 16% des mots recensés par DICOSYN comme synonymes ou antonymes du verbe *protester* se retrouvent dans les voisins de *protester_*CONTRE. Si l'on considère la précision, 3% des voisins sont dans DICOSYN.

Examinons maintenant les 56 mots de DICOSYN qui ne figurent pas dans la liste de voisins, on trouve :

- une importante série de verbes (environ la moitié) qui sont absents du corpus, ou trop rares. Ils n'ont de ce fait pas de voisin distributionnel.

Relation	Nb cooccurrents	Nb voisins
AUPRÈS DE	7	139
AVEC	3	0
CONTRE	392	389
DANS	4	0
DE	10	1
EN	2	0
OBJ	8	24
PAR	2	0
SANS	1	0
SUJ	80	828
À	1	0

TABLE 3.9 – Prédicats associés au verbe *protester*

- des mots qui partagent une partie importante de leurs contextes avec *protester*, mais cette intersection est trop réduite pour passer le seuil de similarité. Cela est dû à une dissymétrie des fréquences entre les prédicats concernés, peu productifs dans le corpus, et le prédicat *protester*_CONTRE, dont la productivité est très importante (392 cooccurrents). C'est par exemple le cas du verbe *s'élever :* une grande proportion des contextes du prédicat *s'élever*_CONTRE (15 sur 21, soit 71%) sont également arguments de *protester*_CONTRE, mais en retour, cette intersection ne concerne que 3% des contextes de *protester*_CONTRE.
- une petite série de verbes qui, tout en étant présents dans le corpus, n'ont pas du tout les mêmes cooccurrents que *protester* , ou partagent avec lui trop peu de cooccurrents (même en considérant d'autres relations que *contre*). C'est le cas par exemple des verbes *accéder, confesser, murmurer.* Les emplois particuliers de ces verbes qui sont considérés comme synonymiques ou antonymiques de *protester* dans DICOSYN ne sont pas représentés dans le corpus.

Ces observations mettent au jour les particularités de l'analyse distributionnelle : fondée sur un corpus, elle en recueille les spécificités, et passe fatalement à côté de mots proches par le sens mais qui ne sont pas utilisés dans ces textes, ou pas dans le sens considéré (cas 1 et 3). En outre, la mesure de similarité ne permet pas de comparer de façon satisfaisante les cooccurrents de deux mots dont les fréquences sont très différentes, même si la mesure de Lin est conçue pour mieux prendre en compte cette situation. Cela reste un cas particulièrement problématique (nous y reviendrons à la section 4.1.2).

Si l'on considère cette fois les voisins qui ne sont pas dans DICOSYN (294 mots), on trouve un certain nombre de mots qui pourraient, à mon sens, venir grossir la liste des synonymes ou des antonymes. Voici une liste de ces propositions complémentaires :

- synonymes (ou para-synonymes[10])
  - que l'on peut gloser par « exprimer un sentiment négatif vis à vis de » : *dé-*

10. Synonymes relevant de deux catégories différentes.

*noncer_*OBJ, *refuser_*OBJ, *critiquer_*OBJ, *déplorer_*OBJ, *condamner_*OBJ, *reprocher_*OBJ, *mettre en garde_*CONTRE, *opposer_*À, *rejeter_*OBJ, *manifestation_*CONTRE, *protestation_*CONTRE, *opposition_*À, *lutte_*CONTRE

  - par extension, que l'on peut gloser par « éprouver un sentiment négatif vis à vis de » : *s'inquiéter_*DE, *craindre_*OBJ, *redouter_*OBJ, *exclure_*OBJ, *regretter_*OBJ, *hostile_*À, *crainte_*DE, *inquiet_*DE

- antonymes (ou para-antonymes)
  - que l'on peut gloser par « exprimer un sentiment positif vis à vis de » : *justifier_*OBJ, *préconiser_*OBJ, *se féliciter_*DE, *plaider_*POUR, *recommander_*OBJ, *demander_*OBJ, *prôner_*OBJ, *partisan_*DE, *accord_*SUR
  - que l'on peut gloser par « éprouver un sentiment positif vis à vis de » : *souhaiter_*OBJ, *vouloir_*OBJ, *exiger_*OBJ, *préférer_*OBJ, *favorable_*À

Le calcul des voisins permettrait donc d'étendre la ressource pour ce mot dans les proportions suivantes, à condition bien sûr de procéder au tri manuel que j'ai été amenée à effectuer :

- 28 % si on considère les 16 mots synonymes et antonymes correspondant à la première glose et sans ajout des voisins intercatégoriels,
- 39 % si on considère les 22 mots après ajout des voisins intercatégoriels,
- 62 % si on prend l'intégralité de la liste des 35 mots.

Ces chiffres n'ont pas de valeur représentative, dans la mesure où le verbe *protester* est probablement un bon candidat au calcul de voisinage sur ce corpus, car il présente des cooccurrents assez spécifiques (la position syntaxique considérée est associée à des restrictions sélectionnelles assez fortes : les mots sont pour l'essentiel des événements). On peut néanmoins voir que la contribution des voisins à une tâche de nature lexicographique pourrait être conséquente, si l'on considère ces résultats comme une ressource assistant un travail manuel de recensement.

Ces cas ne couvrent cependant qu'une petite partie des voisins. Il reste 258 mots qui ne nous semblent pas pouvoir être interprétés selon les relations de synonymie et d'antonymie. On trouve quelques rares cas d'hyperonymie (ou de troponymie selon la terminologie WordNet) que l'on peut gloser par « exprimer un sentiment vis à vis de »: *réagir_*À, *réaction_*À, *porter_*SUR[11], *juger_*OBJ ou par « éprouver un sentiment vis à vis de » : *estimer_*OBJ, *voir_*DANS. Pour le reste, cet exemple illustre bien la diversité des relations que détecte la proximité distributionnelle. Voici la glose de trois autres ensembles de prédicats voisins :

- empêcher la réalisation : *empêcher_*OBJ, *renoncer_*À, *éviter_*OBJ *annulation_*DE, *annuler_*OBJ, *retarder_*OBJ, *interdire_*OBJ, *suspendre_*OBJ, *supprimer_*OBJ, *contrôler_*OBJ, *suspension_*DE
- déclencher la réalisation : *autoriser_*OBJ, *faciliter_*OBJ, *imposer_*OBJ, *ordonner_*OBJ, *procéder_*À, *accélérer_*OBJ, *adopter_*OBJ, *favoriser_*OBJ, *provoquer_*OBJ
- faire état et discuter : *discuter_*DE, *interroger_*SUR, *parler_*DE, *se prononcer_*SUR, *considérer_*OBJ, *évoquer_*OBJ, *envisager_*OBJ, *constater_*OBJ

L'analyse distributionnelle fournit également des relations qui ne relèvent pas de l'axe paradigmatique. Le prédicat *protester_*CONTRE est voisin d'une série de noms comme

11. Il s'agit d'une construction tronquée : *porter sur* X *un jugement*...

modalité, possibilité, rumeur, mesure, vague, série, risque (tous assortis de la relation *de*). (*protester contre une expulsion / possibilité d'une expulsion*).

Cet exemple montre l'intérêt de la base pour la description lexicographique. Il montre aussi le caractère pléthorique de la ressource, ainsi que l'hétérogénéité des relations calculées. A partir de là se pose la question de l'élagage (élimination des voisins non pertinents) et du filtrage (typage des relations calculées) de la base.

3.3.4 Elaguer et filtrer les voisins

Les relations mises au jour sont donc diverses. Comment tirer alors parti d'une ressource de cette nature ? Peut-on filtrer certains liens de voisinage pour identifier une relation particulière (synonymie, antonymie) ? Peut-on éliminer des relations de sens trop lâches ? Selon quels critères et à quelles fins ?

Nous avons cherché dans un premier temps à mieux analyser l'impact des contextes sur la qualité des rapprochements [12]. Les premières observations font apparaître certaines situations problématiques :

- certains mots très fréquents, imposant peu de restrictions sélectionnelles, parasitent le calcul. Ainsi des mots comme *falloir*, *concerner*, *sembler, type*, etc.
- certaines classes de dépendants sont trop homogènes et ne fournissent pas de critère de rapprochement consistant. Ainsi, *banlieue*_DE / *partir*_POUR sont rapprochés par des contextes qui se limitent aux seuls noms de villes.
- certaines relations sont peu discriminantes, comme nous l'avons déjà signalé à propos du tableau 3.9 : c'est en particulier le cas de la relation SUJET ou de certains relations prépositionnelles qui introduisent des compléments circonstanciels, lesquels contraignent faiblement la sémantique du verbe auquel ils s'associent (Ex : *emploi*_EN / *se multiplier*_EN + noms de pays). Ce point est lié au précédent : la relation SUJET est généralement peu fiable car elle rapproche des verbes qui ont pour tout point commun de posséder un complément d'agent humain.
- l'irruption de relations syntagmatiques liées à des alternances très fréquentes de formulation génère des couples d'un type particulier : nous venons de le voir à propos de l'association entre le verbe *protester* et des noms comme *possibilité* ou *série*. De même, des mots comme *genre* ou *type* se retrouvent voisins de nombreux verbes.

Il ne s'agit jamais que d'hypothèses vérifiées de manière sporadique sur des séries d'exemples. On en déduit qu'il est possible de jouer sur plusieurs paramètres pour s'assurer de la validité et de l'interprétabilité d'une relation : la valeur de la mesure de proximité, bien entendu, mais aussi la nature et la diversité des contextes qui servent de base au calcul, ainsi que la fréquence et la variété sémantique des dépendants. Il en découle des pratiques de seuillages de la mesure de proximité, ou d'élimination de relations douteuses, qui sont très difficiles à contrôler. Cette difficulté est due à l'absence d'étalon permettant de mesurer l'amélioration des rapprochements effectués selon les choix opérés : comme les voisins distributionnels excèdent les ressources classiques, il n'y a guère de sens de les utiliser pour l'évaluer. Nous reviendrons dans la troisième partie du mémoire (section 9.1.2)

12. Ce fut l'objet du mémoire de M2 de Ismail El Maarouf, que j'ai co-encadré avec D. Bourigault.

sur ce problème central de l'évaluation.

J'ai fait par conséquent le choix, dans un premier temps, de me focaliser sur un sous-ensemble de la base, et d'aborder ainsi l'étude des voisins à partir d'une étude ponctuelle permettant ainsi un examen plus systématique, comme l'a fait (Tutin, 2007), dont nous reparlerons section 8.2.1. Ce faisant, j'ai poursuivi l'étude des relations intercatégorielles entre noms et verbes.

3.4 Retour sur le lien nom-verbe

Comme nous l'avons dit précédemment, le rapprochement de prédicats donne accès à des relations sémantiques entre catégories de mots différentes. Celles-ci ne se limitent pas aux relations nom-verbe. On trouve des exemples de relations entre verbes et adjectifs, ou entre noms et adjectifs. Ainsi, *hostile*_À est voisin du prédicat verbal *s'opposer*_À et du prédicat nominal *opposition*_À. Je me suis cependant focalisée sur le seul cas du lien nom-verbe, poursuivant ainsi l'exploration initiée dans les travaux mentionnés au chapitre 2. Cela concerne 16 % des couples de voisins impliquant un prédicat nominal. J'ai commencé par le cas des voisins intercatégoriels liés morphologiquement, avant de m'intéresser à tous les types de couples nom-verbe.

3.4.1 Part et nature du lien morphologique

La base des voisins fournit des éléments pour étudier en corpus la nature du lien entre le verbe et le nom qui lui est morphologiquement associé. Ce lien est censé exister en langue, et on fait d'ordinaire l'hypothèse que le verbe et le déverbal processif ont un fonctionnement parallèle, aussi bien sur le plan syntaxique (cadre de sous-catégorisation) que sémantique (Grimshaw, 1992). L'examen des voisins distributionnels montre en réalité une grande disparité de situations, et suggère que ce lien ne va pas de soi si l'on considère les réalisations des deux mots dans un corpus particulier.

On peut observer :

- les cas où le déverbal et le verbe partagent plusieurs relations, confirmant l'hypothèse d'une équivalence sémantique forte. Le voisinage sémantique s'établit sur la base de liens multiples. C'est le cas de couples comme *ouverture* / *ouvrir* ou *ouverture* / *s'ouvrir*, pour lesquels on trouve de nombreuses relations de voisinage [13].
- les cas d'intersection limitée entre dépendants. Le déverbal apparaît comme voisin du verbe, mais avec un score de similarité bas (*jouer*_AVEC/*jeu*_DE).
- les cas où le déverbal entretient une relation privilégiée avec un verbe qui ne correspond pas à sa base verbale (ou vice-versa). Ainsi le verbe *remplir* n'est pas relié à *remplissage* dans le corpus LM10 mais à *respect* (au sens de *remplir une condition*, *une formalité*, etc.), le nom *croissance* est relié plus immédiatement à *progresser* ou *augmenter* qu'à *croître*.

13. Voici l'ensemble des prédicats voisins qui sont détectés autour de ces mots : *ouvrir*_OBJ/*ouverture*_DE, *ouvrir*_À/*ouverture*_À, *ouvrir*_VERS/*ouverture*_VERS, *ouvrir*_SUR/ *ouverture*_SUR, *ouverture*_SUR/*s'ouvrir*_SUR, *s'ouvrir*_À/*ouverture*_À, *s'ouvrir*_SUR/*ouverture*_SUR

Ces observations montrent qu'un nom déverbal n'est pas systématiquement plus proche de son *alter ego* verbal du point de vue morphologique (et vice-versa), mais peut présenter des affinités plus nettes avec un autre verbe, en tous cas dans le corpus considéré. La prise en compte à grande échelle de critères distributionnels permet ainsi de mettre au jour des proximités sémantiques difficilement prédictibles. On replace ainsi la réflexion sur le lien morphologique dans la réalité des emplois en discours. J'ai abordé cette thématique alors que je participais aux travaux du GDR de morphologie[14], ainsi que dans le cadre d'une collaboration avec N. Hathout dans le projet Lexeur.

J'ai ensuite poursuivi l'étude de ces liens de voisinage intercatégoriel, dans le but d'acquérir cette fois des couples nom-verbe sans lien morphologique.

3.4.2 Au-delà du lien morphologique

Dans un travail mené en collaboration avec D. Bourigault (Fabre et Bourigault, 2006), nous avons choisi d'examiner les couples N-V contenus dans les voisins distributionnels de manière à caractériser cette relation sémantique particulière. Nous nous sommes focalisés sur le schéma distributionnel le plus courant, à savoir V_OBJ / Nom_DE (ex : *intéresser*_OBJ / *attention*_DE, *vouloir*_OBJ / *désir*_DE). Ce choix est le résultat d'observations empiriques : les rapprochements sur la base de relations prépositionnelles au niveau du verbe sont marginaux. Seule la relation Nom_DE/V_SUJ est quantitativement équivalente à la relation Nom_DE/V_OBJ, mais elle est nettement plus bruitée. Nous avons introduit un critère de sélection supplémentaire, relatif au rang de voisinage : un couple V-N est retenu si le prédicat verbal apparaît parmi les 10 premiers voisins du prédicat nominal et le prédicat nominal parmi les 100 premiers voisins du prédicat verbal. Là aussi, ce choix est né de l'observation répétée d'une dissymétrie entre les deux éléments du couple : les verbes sont, de façon générale, fortement représentés dans la liste des voisins des noms, ils apparaissent fréquemment parmi les premiers voisins. L'inverse n'est pas vrai. Ces critères ont abouti à la sélection de 3667 couples du corpus LM10.

L'observation de ces couples montre une fois de plus la grande hétérogénéité des associations calculées. On retrouve en particulier le cas des dépendants formant une classe trop restreinte. Ainsi, le couple *carcasse*_DE / *intercepter*_OBJ est lié par les contextes *véhicule*, *camion*, *voiture*, *navire*, *bateau*. Ces deux prédicats n'ont en commun que des éléments de sens très périphériques, liés au fait de pouvoir s'appliquer à des véhicules.

Nous avons alors cherché à tester sur cet échantillon la possibilité d'optimiser la mise au jour d'une relation forte de proximité sémantique, en mettant en œuvre un filtre supplémentaire en aval du calcul distributionnel. Notre démarche a consisté à conforter le lien sémantique en croisant le critère distributionnel avec un filtre contextuel, en d'autres termes à combiner la cooccurrence de second ordre (le voisinage distributionnel) avec la cooccurrence de premier ordre (le voisinage contextuel). L'idée d'exploiter les deux axes pour conforter le repérage d'une relation de nature sémantique n'est pas nouvelle. Le filtrage

14. Ce fut l'objet d'une communication affichée dans le cadre du colloque international sur les noms déverbaux qui s'est tenu à Lille en 2004, intitulée "Affinités sémantiques entre noms et verbes : au-delà du lien morphologique".

par le contexte s'inspire par exemple de la méthode développée par (Lin *et al.*, 2003), qui utilisent des patrons morpho-syntaxiques pour écarter la relation d'antonymie des résultats de l'analyse distributionnelle. Elle est également proche de la méthode de restriction par les cooccurrences mise au point par Hathout et Tanguy (2005) pour valider des informations morphologiques, lesquels s'inspirent eux-mêmes de Xu et Croft (1998). Leur méthode consiste à s'assurer de la bonne formation d'un mot construit en ne retenant que les formes pour lesquelles on trouve le verbe base dans la même page Web. Nous généralisons cette démarche à l'exploration de liens non morphologiques. En outre, l'idée que nous voulons tester est que l'examen des contextes où les deux unités apparaissent ensemble nous permettra de mieux caractériser la nature du lien.

Nous décidons donc de combiner la parenté distributionnelle entre noms et verbes, comme moyen de détecter une parenté sémantique, et la cooccurrence au sein d'une zone textuelle déterminée, comme moyen de conforter la pertinence du lien sémantique et d'aider à sa caractérisation.

Nous imposons pour cela des critères de filtrage sévères : ils consistent à ne conserver que les couples de prédicats qui apparaissent ensemble dans au moins un paragraphe, avec le même dépendant. Ce choix est guidé par plusieurs motivations : le maintien du même dépendant permet de retrouver à l'identique la relation de cooccurrence qui a contribué au rapprochement des deux prédicats ; en stabilisant certains éléments de contexte, la caractérisation de la relation entre les deux prédicats devrait être facilitée ; enfin, nous augmentons les chances de capter des cas relevant de l'équivalence sémantique en contraignant nettement le contexte – la reprise du même dépendant renforce l'hypothèse d'une proximité des deux formulations.

Notre crainte était que le filtre soit trop strict et conduise à réduire considérablement l'ensemble des voisins calculés initialement. Cela ne s'est pas produit, puisque ce filtre retient près de 40 % des couples, soit 1441 voisins. Le tableau 3.10 fournit trois exemples, assortis de leur contexte de cooccurrence. Ils illustrent les cas où la cooccurrence des deux prédicats nominal et verbal dans une zone relativement réduite correspond bien à deux formulations différentes de la même idée.

Le filtrage contextuel une fois appliqué aux couples N-V, il devient possible d'en dresser une typologie indicative, de manière à mesurer l'intérêt de ces relations pour l'acquisition de relations sémantiques. Une première observation générale montre que 31 % des couples sont liés par une relation morphologique, et parmi eux 92 % relèvent de l'association verbe-déverbal processif (*réduction*/*réduire*). Plus de 2/3 des couples ne présentent donc pas de ressemblance formelle. Parmi les 986 couples concernés, nous avons examiné les 200 premiers couples extraits (par ordre décroissant d'occurrences en nombre de paragraphes dans le corpus [15]), en les considérant successivement au sein des dix premiers paragraphes dans lesquels ils cooccurrent.

Cette étude a permis de dégager les principales relations sémantiques instanciées par ces couples :

– 15% sont des (para)-synonymes (*octroi*_DE / *attribuer*_OBJ),

15. De ce fait, il s'agit probablement de résultats optimaux, qui tendent à se dégrader au fur et à mesure que le nombre de contextes partagés décroit. Cela n'a pas été vérifié.

couple de voisins	Exemple
adepte_DE, pratiquer_OBJ + *marche*	Les Finlandais et les Suédois sont en effet de fervents **adeptes de** la **marche** à pied. Ils distancent dans ce domaine leurs voisins européens. Et ce à tout âge. Dans chacun de ces deux pays, 76% et 74% des plus de 55 ans, notamment, **pratiquent** la **marche** au moins une fois tous les quinze jours, contre 40% en moyenne dans l'Union.
raccourcissement_DE, écourter_OBJ + *mandat*	M. Mitterrand n'a pas eu l'occasion, dimanche 12 avril, de donner son avis sur un éventuel **raccourcissement** du **mandat** présidentiel. Il a, en revanche, rappelé qu'élu pour sept ans, c'était à lui seul d'apprécier s'il devait volontairement **écourter** son **mandat** dans le cas où il estimerait n'avoir plus "la capacité" de l'exercer.
image_DE, voir_OBJ + *métier*	Et surtout **l'image de** beaucoup de **métiers** industriels reste peu attractive. [...] Familles et enseignants **voient** encore trop souvent les **métiers** de la production avec les yeux de Zola.

TABLE 3.10 – Exemples de couples V-N cooccurrents

- 15% sont des antonymes (*violation*_DE / *respecter*_OBJ),
- 10% associent actant et activité typique (*responsable*_DE / *diriger*_OBJ),
- 13% désignent des phases successives d'un événement (*signature*_DE, *conclure*_OBJ).

Reste un peu plus de la moitié de relations non pertinentes : erreurs d'étiquetage ou d'analyse syntaxique (19 %), collocations (20 %), relation sémantique trop générale ou non caractérisable (8 %) (exemple : *entrée*_DE et *situer*_OBJ sont voisins parce qu'ils s'appliquent tous deux à des lieux). La part importante de collocations aurait pu être réduite voire éliminée : dans cette première expérience, nous avons fait l'erreur de ne pas contraindre le fait que le dépendant commun corresponde à deux occurrences distinctes du mot – et par conséquent que les deux voisins relèvent bien de deux empans syntaxiques disjoints. Nous avons récupéré de ce fait des cas de cooccurrences liées[16] (exemple : les prédicats *décréter*_OBJ et *levée*_DE sont reliés parce qu'ils apparaissent dans la séquence « levée des sanctions décrétée par », où la relation objet est identifiée à partir de la forme passive).

A l'époque où elle a été menée, cette expérience avait pour objet de vérifier l'intérêt et la possibilité d'identifier des relations N-V sémantiquement pertinentes, et de mieux comprendre la nature de ces rapprochements. Le croisement du voisinage avec des critères de cooccurrence était testé comme une méthode de filtrage de la ressource. Ce travail m'a ensuite amenée à déplacer mon intérêt vers l'examen des zones de cooccurrence ainsi repérées dans les textes, et a ouvert une nouvelle direction de recherche, dont la présentation fait l'objet du prochain chapitre, et qui consiste à croiser voisinage distributionnel et marqueurs

16. Le dispositif a été depuis reprogrammé par Clémentine Adam de manière à supprimer ces cas de cooccurrences liées (cf. section 4.2.3).

contextuels pour le repérage de relations de discours (cf. 4.2.3).

3.5 Perspective applicative : la recherche d'information

Le numéro de la *revue français de linguistique appliquée* (RFLA) coordonné par T. Poibeau et A. Condamines en 2008, intitulé « Linguistique et accès automatisé à l'information : un bilan », regroupe des articles proposant des éléments de réflexion concernant l'apport de la linguistique aux systèmes d'accès à l'information. L'objectif du numéro est à la fois de faire état de cet apport mais également, en retour, d'envisager en quoi ce contexte applicatif peut « interpeller la linguistique et l'amener à ouvrir son champ d'étude » (Poibeau et Condamines, 2008, 6), invitation à laquelle je souscris pleinement. L'article que j'ai proposé avec A. Picton et D. Bourigault pour ce numéro concerne les deux volets de la réflexion. Il fait le bilan d'une expérience d'intégration des voisins distributionnels dans un système de RI, mené dans le cadre du projet TCAN ARIEL, coordonné par J. Mothe et L. Tanguy [17]. L'utilisation de données de ce type était encore inédite dans un système de RI en 2004 (cf. Van der Plas (2008) pour un état de l'art de l'utilisation de techniques distributionnelles basées sur la syntaxe). Nous disposions d'une plate-forme d'expérimentation, RFIEC, développée à l'IRIT autour du moteur de recherche MERCURE (Boughanem *et al.*, 1998), qui nous a permis d'utiliser le protocole d'évaluation CLEF (*Cross Language Evaluation Forum*) et d'enrichir l'indexation des textes à partir des sorties d'UPERY. Le mémoire de M2 d'Aurélie Picton, que j'ai encadré avec D. Bourigault, était consacré à cette expérimentation.

Lorsque nous avons démarré ce travail, une série de travaux antérieurs nous encourageaient dans cette voie. Les pistes de valorisation définies pendant ma thèse dans le contexte de la recherche d'informations à France Télécom (Fabre et Sébillot, 1999) avaient été corroborées par les expériences ultérieures menées à l'IRISA (Claveau *et al.*, 2004). Celles-ci montraient l'impact positif de relations sémantiques diversifiées sur les performances du système utilisé. Les voisins distributionnels permettent donc de tester à plus grande échelle l'hypothèse selon laquelle la tâche d'expansion de requêtes gagnait à s'appuyer sur des liens de proximité sémantique incorporant des relations plus riches que les relations lexicales usuelles, et en particulier des associations intercatégorielles.

Nous nous sommes basés sur une collection de 90 000 textes issus de l'agence télégraphique suisse et du journal *Le Monde* ainsi que sur 50 requêtes assorties des documents pertinents fournis lors de la campagne CLEF 2001. Nous avons exploité le champ titre des requêtes, composé de quelques mots clés, ainsi que le champ narration (<narr>). Voici quelques exemples de requêtes, champ titre :

(2) *ONU / Etats-Unis invasion d'Haïti*
Grèves de la faim
Coupe du monde de football

17. Programme interdisciplinaire CNRS « Traitement des connaissances, apprentissage et NTIC », 2004-2006. Le projet a permis la mise en œuvre de différents niveaux de traitements linguistiques, dont l'étude décrite ici n'est qu'un aspect, et sur lequel nous revenons dans la dernière partie du mémoire, section 8.2.2.

et champ narration associé au premier titre donné en exemple :

(3) *Les documents pertinents relatent aussi bien la discussion sur la décision d'envoyer des Casques bleus américains que l'invasion elle-même. Ils informent aussi sur les conséquences immédiates de l'invasion*

L'expérience a consisté à étendre les mots de la requête avec les voisins calculés par le programme UPERY sur le corpus LM10. Ce corpus présente une intersection avec les textes de la base à traiter, et cette relative compatibilité thématique et temporelle constitue un paramètre *a priori* favorable. Nous nous sommes conformés aux caractéristiques du protocole d'indexation des textes, qui nous a conduit à manipuler des index atomiques (et à renoncer par conséquent à exploiter les termes complexes produits par UPERY) et à leur appliquer une procédure de troncation. La mesure de la performance consiste à estimer la pertinence des documents ramenés dans les premières positions (précision à 5, 10, 15 et 20 documents).

Plusieurs scénarios d'expansion ont été mis en œuvre, de façon à tester différents types de filtres permettant de limiter au mieux le nombre des voisins ajoutés. En effet, une expansion non contrôlée dégrade fortement les performances, puisque plusieurs dizaines de mots supplémentaires peuvent être ajoutés à partir d'une requête titre. Les différents scénarios mis au point sont les suivants :

- Expansion brute (tous les voisins, jaccard $>$ 0,1) ;
- Expansion ciblant certaines catégories : les noms seulement, les verbes seulement ;
- Expansion par sélection manuelle des voisins ;
- *Distributional Feedback* (*DF*) : cette nouvelle mesure, proposée par A. Picton, est une adaptation du *Blind Feedback*, filtre classiquement utilisé en RI. Celui-ci consiste à réinjecter une partie du lexique des textes ramenés par le système pour réaliser l'expansion de la requête. Le DF, quant à lui, propose une modalité nouvelle de sélection du lexique, en utilisant pour cette phase d'expansion les voisins des termes de la requête qui sont présents dans les textes ramenés par le système.

Voici un exemple d'expansion de requête :

Requête titre : *Grèves de la faim*
Expansion non contrôlée : *salarial réclamer création accord sécurité violence soif*
Expansion par DF : *réclamer accord*

On voit sur cet exemple l'intérêt de filtrer les ressources externes à l'aide du vocabulaire des textes concernés. L'expansion de *faim* par *soif* n'a par exemple pas de pertinence. En outre, ce filtre diminue considérablement le nombre de voisins, et permet d'utiliser cette procédure d'expansion sur les requêtes plus longues (le champ <narr>).

Les résultats que nous obtenons sont détaillés dans Picton *et al.* (2008). Les performances globales sont décevantes. Elles sont mauvaises sur toutes les requêtes titres, puisque quel que soit le scénario considéré [18], l'expansion dégrade tous les résultats, ou au mieux les laissent inchangés si l'on considère la précision du DF à 5 documents. Le *blind feedback* est à chaque fois légèrement supérieur à la méthode de base. Une piste plus prometteuse

18. Dans l'ordre décroissant des performances : expansion brute, noms seuls, sélection manuelle, verbes seuls, DF, méthode de base, BF.

concerne néanmoins les requêtes développées, pour lesquelles les chiffres indiquent une légère amélioration de la précision avec le scénario du DF (cf. tableau 3.11), alors que le *blind feedback* a de moins bonnes performances, ce qui pourrait s'expliquer par un comportement différent de cette procédure sur des requêtes longues. Néanmoins, comme le montre le tableau 3.11, cette amélioration n'est pas significative : la précision du résultat concernant les 5, 10, 15 ou 20 premiers documents ramenés par le système s'améliore au mieux de 4 à 5%.

Requêtes développées	P5	P10	P15	P20	Moyenne
Méthode de base	0,3830	0,3106	0,2539	0,2117	0,2898
Blind Feedback	0,3708	0,2958	0,2486	0,2083	0,2809
Distributional Feedback	0,3915	0,3234	0,2539	0,2149	0,2959

TABLE 3.11 – Résultats de l'expansion par les voisins

L'objectif principal du projet ARIEL était d'assortir l'évaluation globale des différents traitements linguistiques réalisés sur la plate-forme d'une étude précise des phénomènes locaux qui, dans les textes, sont susceptibles d'avoir un impact sur les performances du système. Cette analyse plus locale des résultats s'imposait effectivement, dans la mesure où les requêtes se comportent de façon très hétérogène : les résultats après expansion peuvent être nettement dégradés ou nettement améliorés selon les cas. Ce type d'expérience nous a permis de tirer un certain nombre d'indications plus précises, concernant en particulier les classes de mots pour lesquelles il vaut mieux décider de bloquer l'expansion – typiquement les adjectifs ou les noms propres (entités nommées). Par ailleurs, l'étude des requêtes dont le traitement est amélioré par l'expansion montre l'intérêt de disposer d'une ressource qui exploite des relations sémantiques lâches, qu'il s'agisse d'un lien intercatégoriel (*grève*, *réclamer*) ou de liens entre deux noms (*embargo*, *isolement*) qui relèvent d'une relation de type cause-effet.

La méthode du *distributional feedback* se distingue donc par ses potentialités, offrant une amélioration encore non significative, mais encourageante si l'on entre dans l'examen détaillé des requêtes. Son intérêt tient à sa capacité à contrôler l'apport de liens sémantiques larges et diversifiés par un filtrage réalisé à partir des textes de la collection.

3.6 Conclusion

La richesse des informations disponibles dans les bases de voisins distributionnels calculées par le programme UPERY sur de grands corpus est patente. On dispose d'une méthode inductive pour acquérir à grande échelle des indices de proximité sémantique, en s'appuyant sur le principe fondamental de substituabilité. Cette approche intéresse donc aussi bien la linguistique que le TAL. Elle fournit des éléments de description diversifiés sur les mots d'un corpus (affinités de premier et de second ordre) ; le classement des prédicats, en rapprochant des structures et non seulement des mots, offre une manière inédite d'observer les liens de proximité sémantique. Le transfert de l'approche distributionnelle

depuis le contexte favorable des sous-langages, auquel Harris la destinait de manière exclusive pour des raisons théoriques bien précises, vers l'analyse de collections de textes aux caractéristiques linguistiques moins bien définies, aux contours flous, nous a semblé une gageure. Aujourd'hui, ce déplacement ne semble pas vraiment perçu comme problématique dans les autres travaux qui ont implémenté la démarche harrissienne – à supposer qu'il soit même perçu – puisqu'il n'est jamais signalé. De manière tout à fait implicite, on semble faire l'hypothèse que la masse de données maintenant accessibles compense le manque de redondance et de régularité qui pourrait compromettre le calcul de similarité. On a néanmoins observé à plusieurs reprises les effets du traitement de corpus non spécialisés, concernant en particulier le caractère peu discriminant de certains types de contexte, l'ambiguïté de certaines relations, la polysémie des voisins.

Nous avons présenté dans ce chapitre plusieurs directions de recherche, dont certaines ont commencé à être explorées (acquisition de liens non classiques, application à la RI) et d'autres demandent à l'être (acquisition de paraphrases). Des études plus approfondies peuvent être également entreprises sur des types de contextes particuliers (suggérés section 3.3), ou sur la description d'un vocabulaire spécifique (comme l'a fait Tutin (2007) pour la constitution de classes de noms transdisciplinaires des écrits scientifiques). Mais dans tous les cas, nous sommes confrontés à la difficulté d'exploiter une ressource pléthorique et bruitée, qui demande un examen manuel minutieux pour tenter de contenir l'hétérogénéité des liens.

Deux expériences suggèrent une autre piste : le croisement de l'analyse distributionnelle et de l'information contextuelle. Le travail mené autour du lien V-N et les premiers résultats obtenus avec la mesure du *Distributional Feedback* suggèrent de poursuivre en ce sens, et de replonger les voisins dans le texte, autrement dit, d'articuler le lexique et le discours. Cette piste est actuellement explorée dans le cadre d'un projet (VOILADIS) et de deux thèses. Le chapitre suivant présente cette démarche et fait état des premiers résultats.

Chapitre 4

Du lexique au discours

Ce chapitre présente une orientation de recherche que j'ai commencé à explorer depuis 2008, dans le cadre de deux projets de recherche (projet ANR ANNODIS et projet PRES VOILADIS) et par le biais de l'encadrement de deux thèses[1]. Elle a consisté à quitter le terrain des seules relations lexico-syntaxiques pour aborder la dimension du discours.

Dans un premier temps, il s'est agi, comme nous l'avons vu dans le chapitre précédent, de recontextualiser les voisins distributionnels de manière à découvrir dans le corpus d'où ils ont émergé des indices supplémentaires permettant de conforter et de qualifier la relation qui les unit. C'est le sens du travail entamé à partir des liens V-N (section 3.4.2), qui se poursuit par l'encadrement de la thèse de François Morlane-Hondère, à travers une première expérience menée sur la détection des antonymes (section 4.1.2). Dans un deuxième temps, cette démarche de recontextualisation m'a amené à faire l'hypothèse que, ce faisant, les liens sémantiques calculés par l'analyse distributionnelle pouvaient être utilisés pour repérer d'autres phénomènes, de nature discursive cette fois, dont ils pourraient constituer l'amorce (section 4.2.3). Nous explorons, dans le cadre de la thèse de Clémentine Adam, en collaboration avec Philippe Muller (IRIT), les possibilités d'utiliser les voisins distributionnels comme indices de cohésion lexicale, la cohésion étant elle-même le point de départ d'études sur la structure discursive. Dans les deux cas, je parle d'une ouverture de mes travaux sur la dimension discursive, mais il est important de préciser ce j'entends alors par *discours* :

1. le discours qui se déploie dans un corpus de textes particulier, permettant d'accéder à des usages langagiers relevant de conditions de production relativement homogènes. Les corpus sur lesquels j'ai travaillé – *Le Monde*, *Frantext*, *Wikipedia* – sont des instances particulières de genres de discours présentant des caractéristiques communes très générales – discours journalistique, littéraire, encyclopédique. Dans ce cas, je ne considère pas le niveau des textes individuels, mais l'espace global du corpus qui donne accès à l'observation des contextes dans lesquels les mots apparaissent. Je

1. Thèse de doctorat de Clémentine Adam, financée dans le cadre du PRES (projet VOILADIS), 2008-2011, co-encadrée avec Philippe Muller, et co-dirigée par M.-P. Péry-Woodley et Nicholas Asher. Thèse de doctorat de François Morlane-Hondère, financée par un contrat doctoral, 2009-2012, directrice de recherche M.-P. Péry-Woodley.

poursuis en ce sens la démarche d'analyse endogène de D. Bourigault, qui envisageait le corpus que l'analyseur syntaxique traite comme un espace dans lequel il est possible de prélever des indices de désambiguïsation.

2. le discours qui se réalise en textes considérés cette fois comme des unités d'analyse propres (Péry-Woodley, 2000). Dans le cadre du projet VOILADIS je m'intéresse à l'organisation des textes en explorant la possibilité d'utiliser des indices lexicaux pour assister le repérage d'éléments de structure, aussi bien au niveau local (relations de discours) qu'au niveau global (segmentation des textes en blocs homogènes, sur le plan thématique ou fonctionnel).

L'objectif se déplace progressivement : d'un simple moyen pour filtrer et typer les voisins (4.1), le corpus devient lui-même un objet que l'on cherche à caractériser (4.2).

4.1 Croiser indices contextuels et distributionnels

4.1.1 L'hypothèse

Mon objectif était de poursuivre et de systématiser l'approche initiée à l'occasion du travail sur le lien intercatégoriel V-N, à savoir utiliser l'espace du corpus pour chercher à filtrer et typer les voisins, en croisant l'approche distributionnelle avec la recherche d'autres indices contextuels. Dans le cas du lien V-N, j'avais cherché à sélectionner les couples dont les éléments se trouvaient à proximité dans le texte, cette proximité contextuelle pouvant être un indice supplémentaire de proximité sémantique. Le croisement de ces deux plans avait déjà été expérimenté avec succès par Lin *et al.* (2003) : leur but était d'utiliser comme filtres des patrons marquant l'incompatibilité (*from X to Y*, *either X or Y*) pour éliminer les antonymes parmi les couples de mots rapprochés par l'analyse distributionnelle. De même, Mirkin *et al.* (2006), s'intéressant cette fois aux relations d'inférence (ex : *drug therapy* $\rightarrow$ *treatment*) comparent les effets de ces deux méthodes, considérées comme deux paradigmes distincts et complémentaires en termes de méthodes d'acquisition de relations lexicales. C'est dans le prolongement de ces travaux que s'est élaboré le projet de thèse de F. Morlane-Hondère, commencée en octobre 2009. J'évoque ici les premiers développements de ce travail.

4.1.2 Premières expériences sur la relation d'antonymie

Parmi les relations sémantiques que le voisinage distributionnel permet d'acquérir, la relation d'antonymie est celle qui a d'abord retenu notre attention, car elle semble se prêter particulièrement bien au type d'investigation que nous cherchons à mener en aval des voisins. Le travail que je présente ici s'est focalisé sur le cas de l'antonymie adjectivale.

Des paires antonymiques figurent bien sûr parmi les voisins distributionnels, détectées au même titre que la synonymie par le critère de substituabilité. Le tableau 4.1 montre pour quelques adjectifs courants le rang (selon la valeur décroissante de la mesure similarité) du

premier antonyme trouvé dans la liste des voisins [2]. On voit que pour certains adjectifs, l'antonyme figure parmi les tous premiers voisins (il en est ainsi de *moderne/classique*, *partiel/complet*, *riche/pauvre*), alors que pour d'autres (*simple/complexe*, *utile/inutile*) , la part de contextes partagés est nettement moindre.

	premier antonyme trouvé	rang	jaccard
courageux	*timide*	41	0.22
moderne	*classique*	2	0.4
partiel	*complet*	1	0.3
rapide	*lent*	9	0.29
riche	*pauvre*	1	0.32
similaire	*différent*	33	0.28
simple	*complexe*	55	0.23
utile	*inutile*	58	0.18

TABLE 4.1 – Exemple de paires antonymiques dans les voisins distributionnels

Notre objectif a consisté à filtrer les voisins pour essayer de repérer les couples antonymiques en leur sein. L'idée est donc de combiner l'approche distributionnelle avec une approche visant à repérer des patrons morpho-syntaxiques marquant une relation sémantique particulière, exactement dans l'esprit de (Lin *et al.*, 2003).

La relation d'antonymie offre un bon point de départ pour ce type d'investigation. Elle est en effet décrite par plusieurs auteurs, et particulièrement par Jones (2002) et Murphy (2006), approfondissant les observations de Fellbaum (1995) et Justeson et Katz (1991), comme une relation qui serait à la fois paradigmatique et syntagmatique. Rappelons (cf. section 1.3.1) que C. Fellbaum a montré que les antonymes ont une forte propension à apparaître ensemble, et même à marquer une opposition d'autant plus forte qu'ils ont tendance à apparaître conjointement de façon significative. Jones et Murphy ont confirmé cette tendance en montrant que la cooccurrence s'incarne dans des patrons antonymiques réguliers (ex : *ni X ni Y*, *X tout comme Y*, etc.), et que ces apparitions conjointes sont liées à des relations de discours particulières. Murphy a amorcé une réflexion sur le statut constructionnel, au sens de Goldberg, de ces patrons antonymiques (Murphy, 2006).

Le mémoire de Master 1 de François Morlane-Hondère (Morlane-Hondère, 2008) avait permis de mettre en place le dispositif d'analyse de la relation d'antonymie, combinant l'approche distributionnelle et le repérage de patrons antonymiques pour filtrer les voisins. Ceux-ci sont projetés sur le corpus et filtrés par des patrons dédiés au repérage de la relation d'antonymie. Au cours de la première année de thèse, ce dispositif a été repris et évalué. Ce faisant, nous avons élargi notre objectif de départ (trouver les antonymes parmi les voisins) et cherché à utiliser cet outillage dans le but de tester l'hypothèse selon

2. Je m'appuie ici sur la liste d'antonymes fournie par DICOSYN pour décider si les deux mots sont antonymes.

laquelle la relation d'antonymie fonctionnerait à la fois sur les plans syntagmatique et paradigmatique – autrement dit, deux antonymes auraient la double caractéristique d'être substituables dans les mêmes contextes, et d'apparaître de façon récurrente l'un avec l'autre dans des motifs particuliers.

4.1.2.1 L'hypothèse : croiser voisinage distributionnel et patrons antonymiques

Les travaux de Jones comme ceux de Murphy qui sont consacrés à l'étude de l'antonymie en corpus s'appuient sur des listes déjà constituées d'antonymes pour mettre en évidence leurs contextes d'apparition. Il s'agit d'étudier la façon dont des antonymes considérés comme canoniques apparaissent conjointement dans les textes, comme supports de relations de discours particulières. Tout récemment, (Lobanova *et al.*, 2010) se sont appuyés sur des paires antonymiques amorces (*seed antonym pairs*) pour détecter en corpus les patrons porteurs de la relation d'antonymie et découvrir par itération de nouvelles paires d'antonymes, en répliquant la démarche initiée par (Hearst, 1992) sur les hyponymes. L'expérience de Lobanova et de ses collègues démontre la faisabilité d'une approche par patrons pour détecter de nouveaux antonymes. Elle en montre également les faiblesses : pour découvrir de bons antonymes, il faut sélectionner des patrons très spécifiques, et le nombre de couples découverts s'en trouve réduit. Par ailleurs, le taux de précision est relativement bas (entre 16 % et 27 % des paires de mots sont jugées véritablement antonymiques).

Dans le prolongement de ces travaux menés en corpus sur l'antonymie, notre objectif consiste à mesurer l'apport des méthodes d'analyse distributionnelle. Nous combinons donc l'approche par patrons, qui permet de détecter les cas de coprésence des antonymes, et l'approche distributionnelle, qui teste leur tendance à la substituabilité. Notre hypothèse est que la combinaison des deux méthodes devrait fournir la configuration optimale pour la détection des antonymes. On peut argumenter en faveur de cette hypothèse de deux manières :

- les patrons antonymiques fournissent un filtre pour ne retenir des voisins que ceux qui entrent dans des contextes contrastifs,
- le critère de voisinage distributionnel permet de retenir, dans l'ensemble des couples appariés par les patrons antonymiques, ceux qui ont tendance à se substituer l'un à l'autre et sont donc sémantiquement plus proches – ce qui orienterait vers un jugement de canonicité et permettrait de compenser l'effet des patrons déclenchant une interprétation contrastive occasionnelle en discours.

4.1.2.2 Expérimentation et premiers résultats

L'étude réalisée par F. Morlane-Hondère sur le corpus WIKIPEDIA a consisté à tester cette hypothèse. Quatre ensembles de paires d'adjectifs ont été constitués :

1. Ensemble **V** : les voisins distributionnels calculés à partir du corpus. Le seuil de la valeur de Lin étant fixé à 0,1, nous obtenons 275 000 couples.
2. Ensemble **P** : les paires de mots réunies par les patrons antonymiques adaptés de (Jones, 2002) (tableau 4.2). Du fait de la prédominance du patron ambigu X ou

Y (qui peut également relier des synonymes) et de la bonne précision des relations minoritaires, nous avons décidé de nous limiter à un sous-ensemble de couples : ceux qui sont reliés par au moins deux patrons de type distinct, soit 907 couples différents (sur les 13 295 couples adjectivaux extraits au total par la méthode des patrons).

3. Ensemble **P+V+** : les paires de mots qui sont à l'intersection des deux ensembles précédents (612, soit 67% des couples trouvés avec les patrons).
4. Ensemble **P+V−** : les paires de mots qui sont trouvées par application des patrons, mais ne sont pas des voisins distributionnels (295 couples).

patron	exemple
X ou Y	Cette connexion peut être **temporaire** *ou* **définitive**.
à la fois X et Y	Sa production est *à la fois* **fermière** *et* **industrielle**.
entre X et Y	les différences *entre* les vins **blancs** *et* les vins **rouges**
plus/plutôt/moins/autant/aussi (bien) X que Y	Il se déguste *aussi bien* **chaud** *que* **froid**.
X plutôt que Y	Il décrit une Terre **sphérique** *plutôt que* **plate**.
soit X soit Y	Les coups francs sont *soit* **directs** *soit* **indirects**.
ni X ni Y	Il n'est donc *ni* **explicite** *ni* **implicite**.

TABLE 4.2 – Patrons antonymiques

Nous avons ensuite comparé la proportion d'antonymes trouvés dans l'ensemble **P+V+** et dans l'ensemble **P+V-**, en sélectionnant aléatoirement 100 paires dans les deux ensembles. Ce calcul s'appuie sur deux types d'évaluation : le recours à une référence, la partie antonymes du dictionnaire DICOSYN du CRISCO (cf. section 3.2.3) et le recours au jugement de locuteurs.

Les résultats de la confrontation avec le dictionnaire sont présentés dans le tableau 4.3. Ils montrent que la proportion d'antonymes est plus forte lorsque les deux critères se croisent : 32 % des couples qui cooccurrent dans les patrons considérés et sont voisins distributionnels sont listés dans DICOSYN, contre 27 % lorsque ces couples ne sont pas voisins. Mais cette différence n'est pas significative. Notons que cette valeur de précision est nettement meilleure que celle de (Lobanova *et al.*, 2010), qui ne retrouvent que 1% à 3% des antonymes recensés dans les ressources lexicales qu'ils utilisent pour le hollandais. Mais cela s'explique avant tout par le fait que nous disposons d'une ressource lexicographique bien plus complète.

	antonyme dans DICOSYN	synonyme dans DICOSYN	absente de DICOSYN
P+V-	27%	6%	67%
P+V+	32%	4%	64%

TABLE 4.3 – Répartition des paires par rapport à la ressource de référence

Le recours à des questionnaires nous a permis de compléter cette première évaluation en prenant en compte des éléments plus précis de qualification de la relation. Six locuteurs ont été appelés à se prononcer sur la nature de la relation qui unit les paires de mots qui leur sont soumises, en choisissant entre les options suivantes : opposition forte, opposition partielle, synonymie, autre relation sémantique, aucune relation sémantique, ne sais pas. Le questionnaire a porté sur 3 jeux de 100 paires extraites, de manière à ce que 50 paires appartiennent à P+V+, et 50 à P+V-, en faisant une extraction aléatoire. Nous avons sollicité deux juges par jeu de données de manière à pouvoir évaluer le taux d'accord. Sur les trois jeux, le kappa moyen pour la relation d'opposition forte est de 0.64 alors qu'il est de 0.27 pour l'opposition faible. Par conséquent, nous n'avons retenu dans un premier temps que les résultats obtenus pour le jugement d'opposition forte. Ils sont présentés dans le tableau 4.4. Le pourcentage correspond aux paires qui ont été qualifiées de manière identique par les deux juges (vote unanime).

	jeu 1	jeu 2	jeu 3
P+V-	12%	16%	14%
P+V+	11%	16%	16%

TABLE 4.4 – Pourcentage d'antonymes (opposition forte) parmi les paires validées par les juges

Dans cette deuxième expérience d'évaluation, le jugement d'opposition forte est moins important lorsque les paires sont à la fois rapportées par le voisinage distributionnel et par les patrons morpho-syntaxiques, mais cette différence n'est pas, là non plus, significative.

Ces résultats démentent notre hypothèse initiale : dans les deux cas, le jugement d'antonymie n'est pas plus fort sur l'ensemble P+V+ que sur l'ensemble P+V-. Autrement dit, la propension des antonymes à la substituabilité n'est pas un indice permettant de renforcer le jugement d'antonymie (et en particulier, de passer d'un jugement d'opposition faible à un jugement d'opposition forte – jugement de canonicité). Ce résultat, contre-intuitif, ouvre plusieurs pistes de réflexion. Tout d'abord, il suggère une primauté du principe de co-apparition sur le principe de substituabilité. Ce n'est que l'amorce d'une hypothèse qu'il faut envisager avec une grande précaution. En observant les couples de l'ensemble P+V- qui reçoivent un jugement d'opposition forte, on constate que le fait qu'ils ne soient pas voisins distributionnels recouvre en fait plusieurs situations :

- les mots apparaissent très rarement dans le corpus (*dioïque* / *monoïque*). Il est normal que l'analyse distributionnelle ne les détecte pas. Dans ce cas, le fait de les trouver malgré tout dans plusieurs patrons antonymiques renforce le verdict d'antonymie.
- deux antonymes peuvent s'opposer lorsqu'ils qualifient un ensemble très limité de noms. C'est notamment le cas de *annuel* et *vivace*, qui ne s'opposent que lorsqu'ils portent sur les noms *plante* ou *espèce*, ou de *ras* et *long*, qui modifient *poil*. Le fait qu'ils ne soient pas pris en compte par le voisinage distributionnel signale qu'ils partagent des contextes très peu diversifiés.
- le couple d'antonymes présente une dissymétrie nette en termes de fréquence. C'est le cas de couples comme *célèbre* (34 001 occ.) et *anonyme* (2734 occ.), *utile* et *néfaste*, *mauvais* et *bénéfique*. Cette dissymétrie est manifestement compatible avec le jugement d'opposition forte, mais elle empêche la détection par l'AD (cf. section 3.3.3).
- quelques couples présentent les caractéristiques requises pour être détectés par l'analyse distributionnelle (fréquence haute et comparable pour les deux éléments du couple), mais ne le sont pourtant pas. Dans ce cas de figure, il semble que l'on ait affaire à des couples pour lesquels la dimension syntagmatique prime sur le principe de substituabilité. C'est le cas de *coupable* / *innocent*, *gai* / *triste, épais* / *liquide*. Le fonctionnement de ces paires dans le corpus doit être étudié de plus près.

Cette première expérience semble corroborer les hypothèses en faveur d'un fonctionnement fortement syntagmatique de la relation d'antonymie. Elle fournit en outre les moyens de préciser cette thèse car nous disposons désormais d'outils permettant d'observer au niveau des paires d'antonymes particulières leur tendance à manifester les deux types de proximité – constructionnelle et distributionnelle – ou à privilégier la dimension syntagmatique.

La thèse de F. Morlane-Hondère permettra d'approfondir ce travail. Il s'agira en particulier d'essayer de mieux contrôler l'expérience en tenant compte des fréquences des mots considérés, de manière à savoir si la non-substituabilité est un fait véritablement significatif – ou si elle est liée à une fréquence basse de l'un ou l'autre mot, ou des deux. Ce premier résultat nous encourage par ailleurs à reproduire l'expérience pour d'autres relations lexicales. Murphy dit en effet :

> « (...) unlike for other paradigmatic relations, antonym relations are lexical as well as semantic (...) that is, the pairing of words is not based solely on meaning but also on the association of those particular word forms, indicating that the pairings are learned through exposure to them and stored as lexical knowledge. »(Murphy, 2006, p. 2)

Il serait très intéressant de déterminer si ce double fonctionnement paradigmatique et syntagmatique est effectivement plus net dans le cas de l'antonymie que pour la synonymie et l'hyperonymie.

4.2 Des indices lexicaux pour la mise au jour de structures discursives

La deuxième direction de recherche s'inscrit dans la lignée des travaux linguistiques réalisés autour de la notion de cohésion lexicale. Reprenant et prolongeant les propositions de (Halliday et Hasan, 1976) et Hasan (1984), M. Hoey (1991) montre que les phénomènes de répétition lexicale participent de façon prépondérante à l'élaboration de l'organisation textuelle : « lexis organizes text », « lexical cohesion becomes the dominant mode of creating texture ». Le mécanisme de répétition lexicale se nourrit de relations lexicales de nature diverse : à la répétition stricte du même item lexical s'ajoutent des liens entre mots associés par des relations lexicales (synonymie, hyperonymie), mais également, selon Halliday et Hasan, par des relations de proximité sémantique aux contours mal définis - qu'ils nomment de façon peu éclairante collocations, et que Hoey qualifie de « ragbag of lexical relations, many of which have no readily available name » (*op. cit.*, p. 7). De fait, Halliday et Hasan ne donnent guère d'éléments pour cerner la notion, parlant d'associations typiques entre items lexicaux qui cooccurrent de façon régulière. Hoey cherche à spécifier en partie ce « bric-à-brac » en répertoriant parmi les cas de répétition lexicale des mécanismes de répétition complexe (liens morphologiques, par exemple *economy-economist*), de paraphrase simple (substitution synonymique), de paraphrase complexe (par le biais de relations triangulaires : par exemple *hot* est lié à *cold* par une relation de paraphrase simple, *hot* est lié à *heat* par une relation de répétition complexe, d'où il découle que *cold* est lié à *heat* par une relation de paraphrase complexe à trois termes). Ainsi, qu'ils s'agisse de collocations, de répétition ou de paraphrase complexes, les liens entre mots envisagés dans ces travaux sur la cohésion élargissent nettement le champ des relations sémantiques qu'on a coutume de considérer dans le cadre lexical. Cette hypothèse, on l'a vu (section 1.2), a été corroborée par (Morris et Hirst, 1991).

Notre hypothèse est que la relation de voisinage distributionnel peut être utilisée pour étudier la cohésion lexicale et, partant, l'organisation textuelle. Deux caractéristiques de cette relation nous incitent à travailler dans cette direction :

1. par la diversité des relations qu'elle permet de détecter, elle offre la possibilité d'exploiter des liens de proximité sémantique qui ne se limitent pas aux relations lexicales usuelles. La validité de cette hypothèse est difficile à apprécier *a priori* : on propose ici d'utiliser une ressource qui contient des couples de mots dont les relations, très diverses, ne sont pas entièrement caractérisées, pour capter des relations d'association sémantique elles-mêmes mal spécifiées. Le principe de rapprochement distributionnel semble néanmoins une piste intéressante pour donner corps à cette notion d'association sémantique.
2. parce qu'elle est calculée à partir de corpus de textes, elle permet de prendre en compte dans une certaine mesure la nature bidirectionnelle du lien entre lexique et texte, tel que le définit Hoey : « the text provides the context for the creation and interpretation of lexical relations, just as the lexical relations help create the texture of the text ». (*op. cit.*, p. 8). Au lieu d'utiliser des ressources dites génériques, le recours à l'analyse distributionnelle permet d'utiliser une approche inductive pour

détecter une proximité sémantique qui se construit effectivement au sein du corpus.

L'idée d'utiliser les voisins comme une ressource lexicale susceptible de fournir des indices pour travailler sur l'organisation du discours a donné lieu à l'élaboration du projet VOILADIS qui a permis de financer la thèse de C. Adam. Je fais tout d'abord une brève présentation du projet ANR ANNODIS sur lequel j'ai également travaillé et avec lequel il est étroitement articulé.

4.2.1 Le projet ANNODIS

Le projet VOILADIS, qui me permet d'explorer ce deuxième versant de l'articulation entre lexique et discours, a pris sa source dans le cadre du projet ANR ANNODIS, coordonné par M.-P. Péry-Woodley, qui m'a offert la possibilité de travailler pour la première fois sur la thématique du discours. Le projet rassemble des chercheurs des laboratoires CLLE, IRIT (Toulouse) et GREYC (Caen). Son objectif est la production du premier corpus français annoté sur le plan discursif. Je ne retrace pas en détail, dans ce rapport, les travaux que nous avons réalisés dans ce projet. Ils ont fait l'objet de plusieurs publications, dont (Ho-Dac *et al.*, 2009) sur le modèle et la campagne d'annotation, et (Ho-Dac *et al.*, 2010) pour les premiers résultats issus de l'exploitation du corpus annoté. J'indique seulement ici les objectifs du projet et la méthodologie, à grandes lignes.

Le projet articule deux démarches d'annotation :

- une démarche ascendante, qui analyse le discours en procédant de façon compositionnelle à partir de l'élucidation de relations de discours définies dans le cadre de la SDRT. L'annotation d'un texte s'appuie sur sa segmentation préalable en unités de discours, et consiste à définir les relations qui les unissent : relations coordonnantes ou subordonnantes de contraste, de narration, d'élaboration, etc.
- une démarche descendante (dite aussi « macro »), qui considère le texte du point de vue de son organisation globale, et vise le repérage de zones de continuité et de rupture. L'hypothèse est que l'interprétation locale du texte est guidée par le repérage de blocs textuels, dont les modes d'organisation sont variables, et qui fonctionnent à différents échelons du texte, au sein d'un paragraphe, entre paragraphes au sein d'une même section, entre sections voire à l'échelle de tout le texte.

Je m'inscris dans le sous-projet « descendant ». Nous visons plus particulièrement le repérage de deux types de zones textuelles, zones à unité topicale et structures énumératives (désormais SE). Les SE en particulier constituent un objet qui rend compte de la complexité de l'organisation des textes, car il s'agit d'un *pattern* discursif général qui peut s'instancier à différents niveaux de grain, depuis le niveau intraparagraphique, jusqu'au regroupement de plusieurs sections, chaque titre pouvant figurer un item de l'énumération. L'annotation manuelle de ces structures discursives a été guidée par un processus de pré-marquage de traits linguistiques (marqueurs d'intégration linéaires, cadratifs, indices d'amorces et de clôture, etc.) qui permet de guider l'annotation selon le principe d'un repérage de zones denses en marques portentielles de structuration. L'annotation est réalisée au sein de la plateforme d'annotation Glozz, développée au GREYC. Elle est exploitée au moyen de techniques de fouille de données qui ont permis de mettre au jour une typologie de ces

structures ainsi que des regroupements d'indices fortement associés à chaque type de SE. En travaillant dans le cadre de cette approche descendante, je me suis consacrée avec mes collègues à la définition du modèle d'annotation, à la rédaction du guide, à la mise en place des campagnes d'annotation exploratoire puis opérationnelle, enfin à la phase d'exploitation des annotations réalisées. Je reviendrai sur certains de ces aspects méthodologiques dans la troisième partie.

Le projet donne un cadre à mes recherches sur le lien entre lexique et discours en ouvrant plusieurs pistes d'exploration sur ce thème : les éléments lexicaux interviennent dans les phases de pré-marquage (par exemple dans le repérage de structures parallèles pour identifier des zones énumératives). Mais surtout, l'annotation de zones discursives homogènes est le point de départ de travaux de fouille de texte qui permettent d'étudier la corrélation entre zones textuelles et lexique. C'est cet objectif qui m'a amenée à définir un projet associé, avec certains des collègues de CLLE-ERSS et de l'IRIT impliqués dans le projet ANNODIS.

4.2.2 Le projet VOILADIS

Le projet VOILADIS[3] se fixe pour objectif d'explorer une direction de recherche complémentaire au projet ANNODIS, consistant à s'appuyer sur des indices de nature lexicale pour repérer des zones de continuité ou de rupture discursive. Ce projet finance la thèse de C. Adam, que je co-encadre avec P. Muller (directeurs de thèse : N. Asher et M.-P. Péry-Woodley). Le projet prévoit l'exploration de deux pistes de recherche complémentaires, articulées aux deux approches du discours développées dans le projet ANNODIS :

- **l'approche ascendante** : dans la perspective du repérage de relations entre unités de discours, une première direction de recherche consiste à identifier sur le plan lexical le corollaire de certaines relations discursives (antonymes porteurs de relations de contraste, synonymes porteurs de relations de reformulation, etc).
- **l'approche descendante** : une deuxième direction consiste à repérer au niveau « macro » des zones textuelles denses en éléments lexicaux proches, indicateurs de cohésion lexicale et donc de continuité.

En résumé, il s'agit de mesurer l'apport d'indices de niveau lexical pour aider à la mise au jour de structures discursives à différents niveaux de grain. Le projet fournit une opportunité de tester de façon plus systématique l'apport des relations de voisinage distributionnel au niveau discursif.

4.2.3 Voisinage distributionnel et proximité en discours : premiers résultats

Grâce au travail de C. Adam, nous avons commencé l'exploration des deux versants prévus dans le projet : le niveau des relations discursives, et celui de la segmentation au niveau macro.

3. Projet financé dans le cadre du PRES Toulouse (2008-2011).

4.2.3.1 Mise au jour de relations discursives

C. Adam et M. Vergez-Couret [4] ont cherché à tester l'hypothèse esquissée dans (Fabre et Bourigault, 2006) : la cooccurrence de voisins distributionnels pourrrait, dans certaines conditions, constituer un indice fiable de la présence d'une relation de discours. Dans cet article, nous avions suggéré que la localisation au sein d'un même paragraphe du verbe et de son déverbal munis du même dépendant pourrait permettre de repérer des variantes sémantiques (ex : *adepte de la marche / pratiquer la marche*) qui seraient le support de relations discursives – élaborations, reformulations. L'exemple suivant illustre par exemple le cas où le titre d'un article comporte une expression verbale qui est reprise sous une forme nominale dans la première phrase du corps de l'article (*assouplit son projet / assouplissement du projet*). Cette configuration signale la présence d'une relation d'élaboration.

(1) *CAPITAL BSN assouplit son projet de protection du capital*
Dans un entretien aux Echos du 11 septembre, M. Antoine Riboud, président de BSN, a annoncé un assouplissement de son projet de protection contre une offre publique d'achat (OPA).

C. Adam a tout d'abord mis en place une interface d'annotation pour évaluer cette hypothèse de manière plus extensive. Cette démarche préliminaire a montré que si cette configuration s'assortit très souvent d'une relation discursive entre les deux propositions qui incluent les voisins, elle ne semble pas régulièrement associée à un type particulier de relation. Une deuxième étape a donc consisté à introduire d'autres types de contraintes pour viser la mise au jour d'une relation discursive spécifique, à savoir la relation d'élaboration. Le travail décrit dans (Adam et Vergez-Couret, 2010) montre l'apport de l'indice de voisinage distributionnel en combinaison avec un indice ambigu de l'élaboration, le gérondif. L'expérience consiste à comparer la proportion d'élaborations détectées par la présence du gérondif seul, avec celle que l'on trouve lorsqu'on contraint en plus les verbes des deux propositions (et, dans une deuxième expérience, leurs objets également) à être des voisins distributionnels. L'exemple suivant montre un cas où les verbes sont voisins, ainsi que les objets :

(2) *Les Skrulls* ***élargissent*** *(...) leur* ***empire*** *en* ***englobant*** *dans celui-ci les* ***mondes*** *moins avancés qu'ils rencontrent.*

L'amélioration est significative : la précision de la détection d'une élaboration avec le gérondif seul est de 60 %, de 81 % si les deux verbes sont voisins. Lorsque les verbes et leurs objets sont voisins, on obtient 95 % de précision.

Ce premier travail conforte l'hypothèse initiale. Il doit être étendu de manière à diversifier les marqueurs considérés, et à observer l'impact de l'indice de voisinage pour la détection de relations discursives à plus longue distance.

4. Marianne Vergez-Couret, doctorante au sein de CLLE-ERSS (thèse co-dirigée par Myriam Bras et Marie-Paule Péry-Woodley) travaille sur la relation d'élaboration.

4.2.3.2 Segmentation thématique

Adoptant cette fois une approche descendante (au sens défini dans le projet ANNODIS), nous avons commencé à étudier l'apport de la ressource de voisinage distributionnel sur la tâche de segmentation thématique de textes (Hearst, 1997). Cette tâche présente plusieurs propriétés intéressantes pour notre projet :

- l'articulation entre discours et lexique est claire, puisqu'il s'agit de découper le texte en blocs présentant une unité thématique en se basant sur des indices lexicaux marquant la cohésion lexicale.
- l'apport de ressources lexicales a été démontré par des travaux antérieurs, cette ressource pouvant être extérieure (thésaurus) ou construite de façon endogène par des techniques d'apprentissage.
- c'est une tâche bien balisée, avec des procédures d'évaluation qui permettent de comparer les performances du programme avant et après intégration de ressources. Les conditions d'évaluation ne sont néanmoins pas idéales, dans la mesure où déterminer les lieux de rupture thématique est une tâche très difficile à réaliser, même manuellement (Bestgen et Piérard, 2006). La méthode que nous avons choisie de mettre en œuvre est conforme aux procédures usuelles : elle consiste à chercher à assimiler les changements de section à des ruptures thématiques, et à déterminer si les seuils calculés par le programme s'alignent sur ces éléments.

L'expérience a consisté à projeter les relations de voisinage distributionnel sur le corpus à analyser, en prenant en compte les liens entre couples de mots (noms, adjectifs, verbes) qui s'étendent au-delà de la phrase, et à mettre en œuvre un algorithme de segmentation dont les performances sont évaluées en prenant pour référence les titres de premier niveau comme marques de rupture. Parallèlement, l'utilisation de liens de réitération simple et de liens synonymiques est testée pour fournir deux *baselines*. L'exemple présenté figure 4.1 est extrait de (Adam et Morlane-Hondère, 2009). Il montre la diversité des relations détectées en utilisant le critère de voisinage distributionnel, par exemple les couples *pays/frontière*, *nord/sud-est vallée/plaine, nord/sud-est, grand/nombreux.* Les voisins mettent au jour des liens qu'aucune ressource classique ne permet de fournir.

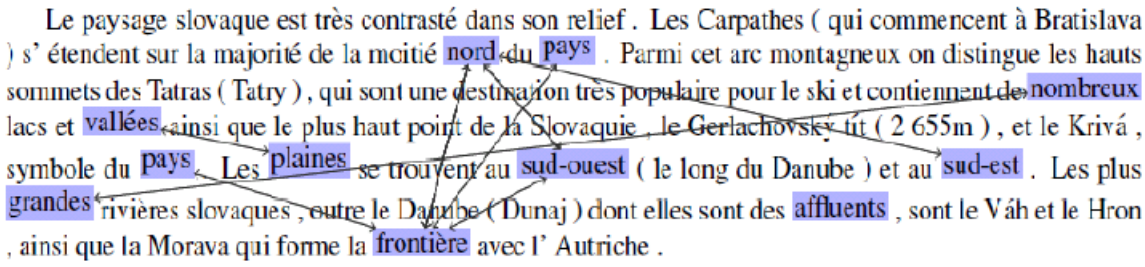

FIGURE 4.1 – Extrait de texte après marquage des liens de voisinage distributionnel

Les premiers résultats obtenus ne sont pas encore concluants en ce qui concerne l'impact des voisins distributionnels sur le traitement. La première expérience menée (*ibid.*) montrait une amélioration des performances par rapport aux deux *baselines* utilisées (exploitation des liens de simple répétition et ajout de synonymes). Ce résultat ne s'est pas confirmé dans la deuxième expérience qui a été menée dans des conditions un peu différentes (en nombre et type de textes considérés) : le calcul basé sur les voisins ne s'est pas, cette fois, avéré meilleur que celui basé sur la répétition stricte. D'autres directions doivent être maintenant explorées. La première consiste à mieux évaluer la faisabilité de la tâche : nous avons montré dans (Adam *et al.*, 2010) l'impact du type de textes sur les performances du segmenteur. Mais surtout, le travail de C. Adam va consister à exploiter les annotations désormais disponibles au terme du projet ANNODIS, aussi bien au niveau ascendant – pour déterminer si des indices lexicaux peuvent améliorer les procédures d'apprentissage pour détecter la présence d'une relation discursive – qu'au niveau descendant – pour examiner les mécanismes de cohésion lexicale qui se manifestent au sein des structures énumératives.

4.3 Conclusion

Notre démarche a consisté à confronter l'approche distributionnelle à une série de tâches auxquelles elle est susceptible de pouvoir contribuer : tâche d'expansion de requêtes en recherche d'information (section 3.5), tâche de segmentation thématique (4.2.2), tâche d'acquisition de relations discursives (l'élaboration) ou sémantiques (l'antonymie) particulières. Dans tous les cas, des traitements réalisés en aval des voisins permettent de les filtrer en s'appuyant sur le corpus traité, pour y exploiter des indices supplémentaires de fréquence et de cooccurrence. Ce faisant, notre objectif est toujours de mieux comprendre la nature de l'information sémantique à laquelle l'AD permet d'accéder. Puisqu'il s'agit d'un indice de substituabilité, dans quelle mesure est-il complémentaire d'approches basées sur l'extraction de patrons ? Puisqu'il fournit des relations sémantiques diversifiées, celles-ci sont-elles un bon moyen de repérer les liens de cohésion lexicale qui tissent le texte ? La définition de dispositifs d'exploration et d'expérimentation adaptés à ces questions est au cœur des travaux en cours.

Deuxième partie

Réexaminer la distinction argument / circonstant à partir des corpus

Présentation de la deuxième partie

Cette partie, consacrée au deuxième volet de mes recherches, concerne la question des affinités syntaxiques entre mots : j'expose les travaux que j'ai réalisés pour développer des outils d'observation du fonctionnement des groupes prépositionnels (désormais GP) dans les corpus. Je me suis intéressée à la traduction en critères mesurables sur corpus de certaines propriétés permettant de déterminer si un GP dépend d'un gouverneur (verbal, ou adjectival) ou s'il a tendance à être autonome. Ces mesures permettent d'examiner avec de nouveaux outils la distinction usuelle entre circonstant et argument.

Le point de départ de cette direction de recherche est ma participation au développement de l'analyseur syntaxique SYNTEX. Le début de la collaboration avec D. Bourigault date du travail sur ZELLIG (3.1), avec l'utilisation des sorties de LEXTER pour fournir au programme d'analyse distributionnelle les contextes syntaxiques permettant de rapprocher les mots d'un corpus (Fabre et Habert, 1998). Pour le type de tâche que nous envisagions, le principe d'une analyse locale, limitée au traitement des SN, s'était avéré inadapté. D. Bourigault évoque cette expérience parmi celles qui, en détournant LEXTER de sa finalité d'origine, ont fait émerger le besoin d'un analyseur syntaxique du français à large couverture et l'ont conduit à lancer ce projet (Bourigault, 2007, p.22). Il m'a contactée peu avant son arrivée à l'ERSS, et nous avons démarré ensemble le projet de développement de SYNTEX en 1999. Un moment impliquée dans le développement proprement dit de SYNTEX ((Bourigault et Fabre, 2000), (Bourigault *et al.*, 2005)), je me suis intéressée ensuite à son utilisation pour l'analyse linguistique, comme outil d'exploration des faits syntaxiques. Je considère alors SYNTEX comme un instrument qui fournit le niveau d'annotation adéquat (étiquetage et dépendances) me permettant d'accéder à certaines régularités morpho-syntaxiques des unités lexicales.

Dans le premier chapitre, je fais brièvement le point sur les différents aspects de ma contribution au projet SYNTEX, puis je présente la façon dont l'analyseur traite le rattachement prépositionnel ; je montre le besoin de passer d'un traitement binaire du GP (il est rattaché à son gouverneur ou il ne l'est pas et, dans ce cas, devient une entité « flottante » dans la phrase) à un traitement qui tienne mieux compte de la fonction qu'occupe le GP par rapport à son gouverneur. Dans les deux chapitres suivants, je présente le travail que j'ai réalisé dans le sens d'une distinction entre différents types de compléments, reprenant ainsi la question de la distinction entre argument et circonstant pour le repérage des compléments du verbe puis de l'adjectif. Dans un premier temps, mon objectif a été de fournir à l'analyseur des éléments permettant d'éviter de rattacher au verbe des GP post-verbaux qui ne présentent pas les propriétés de cohésion requises pour les considérer comme argument. Cette étude m'a amenée ensuite à considérer plus généralement l'apport des sorties de SYNTEX – les liens de dépendance identifiés systématiquement sur un grand volume de données – pour l'étude linguistique des types de complément. Plutôt que de chercher à opérer une distinction entre arguments et ajouts sur la base d'indices statistiques, je me suis intéressée au réexamen de cette distinction, et à la façon dont les données du corpus annoté syntaxiquement offrent des moyens d'observer le continuum souvent envisagé entre ces deux pôles. Comme dans le cas des relations lexicales examinées dans la partie

précédente, j'ai utilisé le dispositif alliant corpus et outils de TAL pour travailler avec d'autres éléments d'observation sur la notion de complémentation, étudiant la façon dont les compléments se comportent dans l'espace du corpus. Le chapitre 6 décrit la méthode, le chapitre 7 fait état des éléments de description linguistique qu'elle contribue à mettre au jour.

Chapitre 5

Le traitement automatique du rattachement prépositionnel dans SYNTEX

Le travail que je présente dans cette partie du mémoire s'inscrit très nettement dans le contexte général de création de l'analyseur SYNTEX, d'abord parce que ce projet émerge au cours de l'activité de développement des modules de rattachement prépositionnel, ensuite parce qu'il illustre un aspect central de l'apport de SYNTEX, son utilisation comme instrument d'observation, pour laquelle plaide D. Bourigault en conclusion de son mémoire d'habilitation :

> « Sans attendre la convergence hypothétique entre une science linguistique à construire et une ingénierie linguistique productrice d'objets techniques, il est possible de mettre en œuvre ces derniers dans une activité de connaissance sur la langue, en les utilisant comme des instruments d'observation (Habert, 2005, chap. VII, et 2006). » (Bourigault, 2007, p. 146)

Dans ce chapitre introductif, j'expose brièvement les différents aspects de ma contribution au projet SYNTEX, avant de présenter ce qui a motivé ma réflexion sur la distinction argument-circonstant, à savoir les choix réalisés dans SYNTEX pour effectuer le rattachement prépositionnel et adjectival dans la phrase et les limites du traitement qui en découle.

5.1 Contribution au projet SYNTEX

Mes travaux autour de SYNTEX ont pris des formes diverses depuis le début du projet lancé par D. Bourigault en 1999.

J'ai brièvement accompagné la phase initiale de développement de l'analyseur en 1999 et 2000, principalement dans la programmation des modules consacrés au traitement des relations ambiguës impliquant le rattachement de groupes prépositionnels (au verbe, au nom, à l'adjectif). Rapidement, ma contribution a quitté le terrain du développement informatique. La mise au point des algorithmes d'analyse s'effectuait par des allers-retours

systématiques entre la conception des heuristiques et l'analyse des résultats sur corpus. Cette démarche expérimentale est au cœur du développement de l'analyseur et m'a amenée à déplacer mon intérêt vers la phase d'analyse des sorties de SYNTEX, et à m'intéresser à certaines problématiques linguistiques spécifiques, une fois les programmes suffisamment stabilisés pour produire des résultats exploitables. C'est ainsi que je me suis focalisée sur l'analyse de la résolution des ambiguités de rattachement prépositionnel qui est au cœur de l'analyse syntaxique automatique. La thèse de doctorat de Cécile Frérot [1], encadrée par Didier Bourigault, est consacrée à cette question et à l'apport de ressources externes à l'analyse, qu'elles soient construites manuellement (comme dans le cas du lexique-grammaire), ou automatiquement (Frérot, 2005). J'ai contribué à cette réflexion dans le cadre d'un travail préliminaire consacré au traitement des compléments verbaux en *de* (Frérot *et al.*, 2003). J'ai entamé parallèlement avec C. Frérot un travail sur la distinction entre types de compléments qui prendra par la suite plus d'ampleur (Fabre et Frérot, 2002).

Une fois SYNTEX stabilisé, le travail sur l'analyseur s'est également accompagné d'efforts pour valoriser son utilisation auprès de la communauté linguistique : j'ai coordonné un projet ILF visant la mise en place d'une collaboration autour de SYNTEX entre des chercheurs linguistes informaticiens, développeurs d'outils d'exploration de corpus (D. Bourigault, L. Tanguy et moi-même), et des linguistes de l'équipe niçoise *Bases, Corpus, Langage* (autour de Sylvie Mellet) possédant une expérience dans le domaine de la description linguistique sur corpus. J'ai participé également au projet ACI d'Elisabeth Delais-Roussarie dans le but de promouvoir l'utilisation de SYNTEX pour étudier de façon extensive les unités syntaxiques dans des corpus oraux. Ces collaborations nous ont amené à intégrer dans nos pratiques des outils d'interface (dans un premier temps TIGERSEARCH [2]) dans le but de permettre aux linguistes partenaires de lancer des requêtes sur des corpus analysés. Elles ont conduit également à l'amélioration de certains modules de traitement, pour analyser des phénomènes qui intéressaient ces chercheurs (meilleure analyse des propositions relatives, traitement des groupes prépositionnels en début de phrase).

Concernant l'exploitation de l'analyseur, rappelons que la motivation initiale de ma collaboration au projet SYNTEX avait été le développement d'un outil permettant de changer d'échelle dans la mise en œuvre des méthodes d'analyse distributionnelle automatique que j'avais expérimentées en travaillant avec B. Habert sur ZELLIG (cf. section 3.1). Dans un deuxième temps, SYNTEX a été utilisé dans un contexte industriel, dans le cadre d'un projet avec le BEA (Bureau d'enquêtes et d'analyses de l'aviation civile) (Hermann *et al.*, 2008), dont l'objectif était de concevoir une chaîne de traitement permettant d'assister la vérification du codage de rapports d'accidents d'avion, ainsi que l'activité de codage proprement dite. Cette collaboration a d'abord pris la forme d'un projet impliquant les étudiants du master ECIL (projet tutoré « conception »), que j'ai animé avec D. Bourigault. SYNTEX est ainsi devenu également un outil précieux sur le plan pédagogique, pour présenter les méthodes et les résultats de l'analyse syntaxique automatique.

1. Cécile Frérot est aujourd'hui maître de conférences à l'Université Stendhal - Grenoble 3.

2. Développé par l'IMS de l'université de Stuttgart. Depuis, Franck Sajous a développé à CLLE-ERSS un concordancier sur les sorties de SYNTEX, INCAS (Interface de Navigation en Corpus Annotés Syntaxiquement).

Le développement d'un outil comme SYNTEX est donc indissociable de sa valorisation. C'est de fait principalement en aval de l'analyseur que se situe ma contribution. Mon objectif principal est devenu d'utiliser les sorties de SYNTEX comme point de départ de l'analyse linguistique. Je cherche à montrer que le fait de disposer de grands volumes de données analysées automatiquement constitue un atout pour l'étude du langage. En clôture de son mémoire d'habilitation, D. Bourigault plaide pour ce type d'utilisation de SYNTEX :

> « (....) c'est d'abord dans l'utilisation de l'analyseur en tant qu'instrument que l'on doit attendre la production de savoirs diversifiés sur la langue. » Bourigault (2007, p.145)

Dans cette perspective, je me suis intéressée à la question du rattachement prépositionnel.

5.2 Procédures de rattachement prépositionnel

SYNTEX est basé sur le principe d'une analyse en dépendance qui établit une relation entre un gouverneur et un dépendant. De proche en proche, l'analyse en dépendance produit des liens qui constituent une clause. Le traitement s'effectue en trois passes :

1. Identification des liens de dépendance locale : en particulier, ce module identifie le lien de la préposition vers son dépendant et analyse les complexes verbaux – séquences telles que *a été vu*, *peut simplement s'être trompé*, etc. Le traitement de la coordination intervient à l'issue de cette passe.
2. Identification des liens de dépendance non ambiguë (sujet, attribut et objet).
3. Résolution des liens de dépendance ambiguë (rattachement des prépositions, des adjectifs et des antécédents de pronoms relatifs).

Le traitement des adjectifs et des groupes prépositionnels s'appuie donc sur la prise en compte d'indices de désambiguïsation.

5.2.1 Principes de désambiguïsation

La phase de désambiguïsation est basée sur des indices endogènes et exogènes. L'approche endogène consiste à apprendre les propriétés lexico-syntaxiques des mots du corpus analysé sur des cas non ambigus et à utiliser cette connaissance pour résoudre les cas d'ambiguïté locale. Cette démarche a été adoptée de façon exclusive au démarrage du projet SYNTEX. L'hypothèse, issue des expériences menées par D. Bourigault avec LEXTER sur des corpus spécialisés dans le champ de l'acquisition de terminologie, était que la projection de lexiques externes était inadéquate car elle ne permettait pas de capter la singularité des fonctionnements syntaxiques dans ces corpus. Cette hypothèse de départ a été revue progressivement au fur et à mesure de l'élargissement du champ d'action de l'analyseur vers le traitement de grands corpus diversifiés. Elle a donné place à une stratégie mixte combinant des procédures endogènes et des ressources exogènes. La thèse de C. Frérot a été consacrée à cette problématique, et plus particulièrement à l'intérêt de recourir à une ressource de type lexique-grammaire. Nous avions commencé à explorer ensemble la faisabilité de cette idée sur le traitement de la préposition *de* (Frérot *et al.*, 2003). Dans

cet article, nous avions montré que la formule la plus efficace pour faire collaborer indices endogènes et ressources externes consistait à exploiter leur intersection plutôt que leur complémentarité : les techniques endogènes permettent de filtrer la ressource lexicale constituée *a priori* et c'est l'appartenance d'un verbe à l'intersection des deux listes qui en fait un excellent candidat au rattachement. Les apports spécifiques à chacune des deux sources d'information s'avèraient moins fiables. Ce résultat relativisait nettement l'apport potentiel d'une ressource externe de type lexique-grammaire, ce que la thèse de C. Frérot a ensuite confirmé. D. Bourigault a ensuite opté avec C. Frérot pour une autre stratégie consistant à construire un lexique comportant des probabilités de sous-catégorisation apprises à partir d'un grand corpus. Cette formule exogène – au sens où les données sont apprises sur un corpus externe et peuvent être réexploitées sur des corpus d'analyse de différents types – s'est révélée la plus efficace.

Ces indices sont utilisés lorsque l'analyseur identifie plusieurs candidats possibles pour régir le groupe qu'il cherche à attacher. En ce qui concerne les GP, SYNTEX recherche les gouverneurs potentiels de chaque préposition (verbe, nom, adjectif) en parcourant de droite à gauche la séquence à analyser, jusqu'à rencontrer certaines configurations d'arrêt (présence d'un verbe, d'un pronom relatif, etc.). Lorsque plusieurs candidats gouverneurs sont identifiés, l'analyse cherche à les départager en exploitant des indices endogènes et/ou exogènes (cf. Bourigault, 2007, 114 et s.).

Les indices endogènes utilisés relèvent principalement de deux types d'information :

la fréquence : pour départager deux gouverneurs possibles w_1 et w_2 pour une préposition p et un mot w' lui-même gouverné par la préposition, on compare la fréquence des 2 triplets correspondants, soit (w_1, p, w') et (w_2, p, w'). Cette valeur de fréquence est obtenue en récoltant ces informations sur des configurations non ambiguës. On tient compte aussi de la fréquence avec laquelle le gouverneur apparaît sans cette préposition.

la productivité : elle est utilisée pour estimer la force d'une relation entre un gouverneur et la préposition. La productivité du couple (w, p) est définie comme le nombre de triplets (w, p, w'), autrement dit le nombre de mots différents qui dépendent de la préposition lorsque celle-ci est rattachée au mot w. On élimine les hapax en ne prenant en compte que les triplets dont la fréquence est au moins égale à 2. Plus précisément :

$$prod(w, p) = Card\{w'/f(w, p, w') \geq 2\}$$

On raisonne alors sur les types, et non plus sur les occurrences. La productivité permet d'estimer le « rendement de[s] configurations syntaxiques » (Legallois, 2005), leur régularité, sur la base de leur capacité à accueillir un nombre important de lexèmes.

5.2.2 Le cas du rattachement au verbe

L'exemple 1 montre un cas de phrase simple analysée par SYNTEX, dont la figure 5.1 offre une représentation graphique simplifiée.

FIGURE 5.1 – Exemples de liens de dépendance posés par SYNTEX

(1) `<TXT>` On n' échappe pas à ce choix .
`<ETIQ>Pro|on|On|1|SUJ;3|  Adv|ne|n'|2|ADV;3|`
`VCONJS|échapper|échappe|3||SUJ;1,ADV;2,ADV;4,PREP;5`
`Adv|pas|pas|4|ADV;3|`
`Prep|à|à|5|PREP;3|NOMPREP;7  DetMS|ce|ce|6|DET;7|`
`NomMS|choix|choix|7|NOMPREP;5|DET;6  Typo|.|.|8||`

Chaque *token* est caractérisé par 6 champs d'information (j'illustre chacun à partir du *token choix*) :

1. sa catégorie (`NomMS`)
2. son lemme (`choix`)
3. sa forme (`choix`)
4. sa position dans la phrase courante (`7`)
5. la relation qui l'unit à son gouverneur et la position de celui-ci (`NOMPREP;5` : le nom *choix* est gouverné par le *token à* par la relation NOMPREP.)
6. la relation qui l'unit à son (ou ses) dépendant(s) ainsi que sa (ses) position(s) (`DET;6` : le nom *choix* gouverne le *token ce* par le biais de la relation DET).

L'exemple 1 montre que le GP est bien relié au verbe : le *token à* figure dans sa liste de dépendants, et il le gouverne par la relation PREP.

Plusieurs aspects du traitement sont à considérer :

– Le gouverneur du GP est recherché exclusivement sur sa gauche. Cette recherche peut remonter assez loin, tant que certains cas de figure ne sont pas rencontrés (présence d'un verbe, ou d'un nom dont le gouverneur se situe à droite, etc.). Le rattachement entre le gouverneur et la préposition peut par conséquent s'effectuer à distance, comme dans l'exemple 2, qui montre que le rattachement peut s'opérer par-dessus un adverbe et un élément en incise (dans les exemples qui suivent, je ne restitue pas toute l'analyse, mais uniquement la notation correspondant au gouverneur verbal et au GP).

(2) `<TXT>` Et la réponse tient sans doute, en partie, à la pugnacité (...)
`<ETIQ> VCONJS|tenir|tient|4||SUJ;3,ADV;5,PREP;9`
`Prep|à|à|9|PREP;4|NOMPREP;11`
`Det??|le|la|10|DET;11|`
`NomFS|pugnacité|pugnacité|11|NOMPREP;9|DET;10,PREP;12`

– Il découle du point précédent que tous les GP préverbaux (l'ensemble des GP dont le gouverneur verbal est situé à leur droite) sont laissés flottants, sans gouverneur, et sont assimilés à une clause. Cette décision a été prise de manière à ne pas prendre le risque de rattacher au verbe des GP qui relèvent du niveau de la phrase. Evidemment, cette décision n'est pas valide pour tous les GP trouvés dans

cette position (cf. 3), mais la démarche semble raisonnable si l'on considère la plus grande partie des GP en initiale de phrase. L'exemple 3 fournit une illustration de ce cas de figure. On voit que la préposition *en* n'est pas gouvernée par le verbe.

(3) `<TXT>` En janvier 1943 , Londres donne son accord .
`<ETIQ> Prep|en|En|1||NOMPREP;2`
`NomXXDate|janvier 1943|janvier 1943|2|NOMPREP;1|`
`VCONJS|donner|donne|5||SUJ;4,OBJ;7`

D'autres positions bloquent également la tentative de rattachement : c'est le cas des GP en incise (ex : *avait donné le coup d'envoi, en septembre,...*).

– SYNTEX rattache un GP à son gouverneur sans déterminer la fonction qui sous-tend cette relation. Dans l'exemple 4 le GP *à l'audience* (circonstant temporel) est rattaché au verbe *découvrir* de la même manière que, dans l'exemple 2, *à la pugnacité* (complément indirect) est rattaché au verbe *tenir*.

(4) `<TXT>` M. Alain Schrantz (...) qui paraît découvrir à l' audience que ...
`<ETIQ> VINF|découvrir|découvrir|18|ATTS;17|PREP;19,OBJ;22`
`Prep|à|à|19|PREP;18|NOMPREP;21`
`Det??|le|l'|20|DET;21|`
`Nom?S|audience|audience|21|NOMPREP;19|DET;20`

L'exemple 5 permet de faire la synthèse des limites du traitement réalisé par SYNTEX : les deux GP *en mars* et *en France* sont rattachés de la même manière au gouverneur verbal *venir*, alors qu'ils n'ont pas la même relation au verbe : le premier exprime une circonstance temporelle, le second un complément de destination requis par la sémantique du verbe. Par ailleurs, placé avant le verbe, le GP *en mars* aurait automatiquement été laissé flottant, et il n'est pas certain dans ce cas précis que la décision de rattachement dépende uniquement de la position par rapport au verbe. On aimerait donc pouvoir affiner ces décisions de rattachement.

(5) `<TXT>` Nous devons venir en mars en France .
`<ETIQ> VINF|venir|venir|3|OBJ;2|PREP;4,PREP;6`
`Prep|en|en|4|PREP;3|NOMPREP;5`
`NomXXDate|mars|mars|5|NOMPREP;4|`
`Prep|en|en|6|PREP;3|NOMPREP;7`
`NomPr|France|France|7|NOMPREP;6|`

5.2.3 Dépendants adjectivaux

Je me suis également intéressée au cas du rattachement des GP à un gouverneur adjectival, et je résume ici les éléments d'analyse auxquels je ferai appel à la section 6.5 (cf. Bourigault, 2007, 120 sq.).

La recherche du mot qui gouverne l'adjectif se déroule cette fois à gauche et à droite de celui-ci. La désambiguïsation du rattachement adjectival est effectuée essentiellement sur la

base de l'indice endogène de fréquence. Ainsi, dans le contexte *plainte pour discrimination déposée*, l'adjectif *déposée* peut être rattaché soit à *plainte*, soit à *discrimination*. Pour guider la décision, on compare la fréquence des configurations non ambiguës dans lesquelles on a rattaché *déposée* à *plainte* ou à *discrimination*[3].

Alors que la préposition est rattachée à son gouverneur verbal par une relation unique (PREP) plusieurs relations peuvent être identifiées dans le cas de l'adjectif :

1. relation attribut du sujet (relation ATTS).

   (6) `<TXT>` Notre détention a été terrible
   `<ETIQ> PpaMS|être|été|5|AUX;4|ATTS;6`
   `Adj??|terrible|terrible|6|ATTS;5|`

2. adjectif épithète (relation ADJ)

   (7) `<TXT>` Les trois frères ... ont reçu des soins médicaux
   `<ETIQ> Adj?P|trois|trois|2|ADJ;3|`
   `Nom?P|frère|frères|3|SUJ;19|DET;1,ADJ;2,REL;5`
   `Nom?P|soin|soins|22|OBJ;20|DET;21,ADJ;23`
   `Adj??|médical|médicaux|23|`

3. attribut de l'objet (relation ATTO)

   (8) `<TXT>` Il s' est rendu tristement célèbre ...
   `<ETIQ> PpaMS|rendre|rendu|4|AUX;3|ATTO;6,PREP;7`
   `Adj??|célèbre|célèbre|6|ATTO;4|ADV;5`

4. apposition (APPOS).

   (9) `<TXT>` les DRH , soucieux de voir leur entreprise remporter le trophée ...
   `<ETIQ> AdjM?|soucieux|soucieux|10|APPOS;8|PREP;11`

Si aucun gouverneur n'est identifié, l'adjectif porte la relation NOGOV. C'est le cas lorsque l'adjectif est en situation d'incise en début de phrase, ou lorsqu'une configuration ambiguë n'a pas pu être résolue. Par exemple, dans :

(10) *L'Inde rend Islamabad responsable du détournement*

responsable peut être analysé comme attribut de l'objet ou comme qualificatif du nom *Islamabad*, et l'analyseur ne dispose pas des informations permettant de prendre une décision fiable, d'où le choix de ne pas trancher.

A titre indicatif, sur le corpus LM10, 30 % des adjectifs sont régis par la relation ATTS, 29 % par la relation ADJ, 11 % par la relation APPOS. La relation ATTO est très rarement identifée (2 %). Près d'un tiers des gouverneurs des adjectifs (29 %) ne sont pas identifiés.

3. Dans le cas des formes étiquetées comme participes passés, les relations calculées pour le verbe correspondant entrent aussi en jeu.

Considéré cette fois comme gouverneur d'un GP, l'adjectif est traité de la même manière que le verbe, qu'il s'agisse, comme dans l'exemple 11, d'un élément sous-catégorisé par l'adjectif :

(11) <TXT> applicable à toutes les chaînes
<ETIQ>
Adj??|applicable|applicable|21|ADJ;20|PREP;22
Prep|à|à|22|PREP;21|NOMPREP;24
Det??|tout le|toutes les|23|DET;24|
NomFP|chaîne|chaînes|24|NOMPREP;22|DET;23

ou, dans l'exemple 12, d'un GP plus périphérique :

(12) <TXT> Mais les syndicats voudraient qu' une seule loi soit applicable dans le pays
<ETIQ> ??|applicable|applicable|10|ATTS;9|PREP;11
Prep|dans|dans|11|PREP;10|NOMPREP;13

5.3 Conclusion

Je me suis intéressée à la possibilité d'aller au-delà de la décision binaire qui consiste à rattacher un GP à un gouverneur ou à le laisser flotter. J'ai ainsi cherché les moyens d'approcher avec des techniques de TAL l'opposition traditionnelle entre arguments et circonstants. Cette démarche a été inspirée par l'approche endogène qui est menée dans SYNTEX, et qui consiste à capter des informations sur l'ensemble du corpus pour guider des décisions de désambiguïsation à un niveau local.

Lorsque j'ai commencé à réfléchir à cette question avec D. Bourigault en 2000, quelques rares travaux antérieurs avaient envisagé de capter automatiquement la distinction entre arguments et ajouts. M. Brent considérait avec prudence la possibilité d'un lien entre fréquence et type de compléments.

> « It is not clear how a machine learning system could do this, although frequency might provide some clue. »(Brent, 1993)

Mais il estimait que la tâche était trop mal spécifiée pour être tentée :

> « Until a useful definition that trained humans can agree on is developed, it would seem fruitless to attempt machine learning experiments in this domain. » (*ibid.*)

Les travaux visant l'acquisition automatique de cadres de sous-catégorisation verbale faisaient plus ou moins explicitement référence à la distinction entre arguments et autres types d'expansion :

> « Simple probabilistic models may be used to denote verb arguments according to the variance of their distributions throughout domains and sublanguages. » (Basili *et al.*, 1997, 165).

Mais aucun de ces programmes ne cherchait à évaluer spécifiquement leur capacité à réaliser cette distinction.

Nous avons fait alors l'hypothèse que le fait de pouvoir examiner les occurrences des GP dans de vastes corpus de textes avec des outils d'exploration adaptés pouvait nous donner accès à de nouvelles méthodes d'expérimentation et d'analyse.

Chapitre 6

Approche *corpus-driven*[1] de la distinction argument / circonstant

Dès le début du développement de SYNTEX (Fabre et Bourigault, 2001), nous nous interrogeons sur le moyen de prendre en compte l'hétérogénéité du lien entre le GP et son gouverneur. Notre objectif est de déterminer s'il est possible d'aller au-delà du simple diagnostic de rattachement, et d'utiliser des indices issus du corpus de traitement pour opérer une distinction entre des groupes sélectionnés par le gouverneur – les arguments, ou compléments – et des éléments satellites plus autonomes – les circonstants, ou ajouts. Même si la terminologie argument/circonstant est plus usuelle, je parlerai plutôt d'une opposition entre argument (ou complément) et ajout, la notion d'ajout étant plus générale (et usitée en anglais) et pouvant rendre compte plus clairement de la diversité des GP qui n'entrent pas dans la valence du verbe, qu'il s'agisse de circonstants proprement dits ou de compléments de phrase, voire d'adverbiaux cadratifs susceptibles d'avoir une portée au-delà de la phrase (Charolles et Péry-Woodley, 2005).

J'ai abordé cette réflexion avec C. Frérot dans (Fabre et Frérot, 2002). D. Bourigault m'a aidé à reprendre et à finaliser cette première version de la méthode (Fabre et Bourigault, 2008). J'ai poursuivi ensuite cette réflexion, sur le plan de la caractérisation linguistique des résultats, avec J. Rebeyrolle. Une collaboration ponctuelle avec M. Ho-Dac nous a permis de réfléchir à l'intégration de critères positionnels (Fabre *et al.*, 2008). Enfin, une collaboration avec Anna Kupść (CLLE-ERSS à Bordeaux) m'a amenée à prolonger ce travail par l'examen des compléments adjectivaux (Fabre et Kupść, 2009).

6.1 Caractérisation linguistique

La linguistique oppose les compléments essentiels du verbe, qui fournissent les arguments du prédicat verbal, aux éléments périphériques – ajouts, circonstants. Selon les termes d'Andrée Borillo on distingue en effet « les compléments qui s'attachent [au verbe] dans des combinaisons bien réglées » et font partie de son « schéma de construction », et les

1. Le terme *corpus-driven* est défini et discuté section 9.2.2.

éléments « qui apparaissent dans son environnement mais qui ne dépendent pas directement de ses propriétés structurelles» (Borillo, 1990). Les deux types de dépendants n'occupent pas les mêmes positions dans la structure de la phrase : les arguments dépendent du verbe, les ajouts peuvent se rattacher à différents nœuds de l'arbre syntaxique.

> « Une partie des dépendants d'une tête fonctionnent comme des arguments : ils sont syntaxiquement et sémantiquement sélectionnés par la tête. D'autres fonctionnent comme des ajouts : ils ne sont pas sélectionnés par la tête, mais sélectionnent au contraire la tête avec laquelle ils se combinent. » Bonami (1999, p. 9).

J. Grimshaw, citée par (Merlo et Ferrer, 2006), oppose des arguments sémantiquement sélectionnés par la tête, et des ajouts qui ont une sémantique stable, non conditionnée par la relation qui les lie à la tête :

> « With respect to their interpretation, a complement is an argument if its interpretation depends exclusively on the head with which it is associated, whereas it is an adjunct if its interpretation remains relatively constant when associating with different heads. » Grimshaw (1992, p. 108)

Une batterie de tests a été définie pour tenter de distinguer les deux types de dépendants.

O. Bonami, considérant l'ensemble des tests, conclut que la distinction est valide, mais qu'elle ne peut être approchée qu'en termes de conditions suffisantes très spécifiques. P. Miller insiste plutôt sur la difficulté d'appliquer les tests, sur le fait qu'ils peuvent amener vers des conclusions contradictoires, et se prononce en faveur d'un continuum, reliant des arguments prototypiques (les arguments directs) et des ajouts prototypiques (les circonstants temporels), avec entre les deux une zone d'indétermination (Miller, 1998). C. Manning présente cette question comme une des meilleures illustrations de l'intérêt de passer d'une vision catégorique à une vision probabiliste de la syntaxe :

> « This conception of the argument/adjunct distinction is the best one can do in the categorical 0/1 world of traditional formal grammars: things have to be either selected (as arguments) or not. (...) However, categorical models of selection have always been problematic. The general problem with this kind of model was noticed early on by Sapir (1921:38) who noted that "All grammars leak". In context, language is used more flexibly than such a model suggests. » (Manning, 2003)

Nous verrons comment la méthode que nous proposons permet de rendre compte de cette flexibilité, et du caractère graduel de cette notion. Notre point de départ consiste cependant à faire le point sur les critères proposés pour rendre compte de cette opposition. Je présente donc tout d'abord les tests que discute O. Bonami dans sa thèse, car j'y ferai appel à plusieurs reprises. O. Bonami propose une discussion approfondie de l'ensemble des critères traditionnellement mis en avant :

Caractère obligatoire au niveau syntaxique Il est bien connu que de nombreux verbes admettent facilement l'omission de leur argument. O. Bonami reformule ce critère comme une condition suffisante, mais non nécessaire :

« si un dépendant est syntaxiquement obligatoire, alors c'est un argument. » (Bonami, 1999, p. 24) [2].

Caractère obligatoire au niveau sémantique Un dépendant peut être syntaxiquement optionnel mais requis sur le plan sémantique : même omis, sa place est occupée et donne lieu à des inférences :

(1) *Paul mange* → *Paul mange quelque chose*

De fait, tous les arguments directs sont sémantiquement obligatoires. Mais de nombreux types de dépendants qui semblent de nature argumentale ne sont pas pour autant requis sémantiquement. C'est le cas des datifs :

(2) *Paul chante* ↛ *Paul chante à quelqu'un.*

Par ailleurs, de nombreuses classes d'ajouts sont sémantiquement obligatoires (ajouts temporels et locatifs, ajouts de manière et de cause pour les verbes d'action), ce qui n'en fait pas non plus une condition suffisante.

Itérabilité Selon ce critère, plusieurs instances du même type d'ajout peuvent dépendre de la même tête verbale, ce qui n'est pas le cas des arguments. O. Bonami montre qu'en réalité, l'itérabilité est une propriété des expressions temporelles et spatiales, et qu'elle est maintenue lorsque celles-ci ont un statut argumental :

(3) *Jean est allé dans le parc au bord de l'étang près du saule.*

D'autre part, cette propriété s'applique de façon limitée aux ajouts.

Positions syntaxiques Les arguments et les ajouts se distinguent par la nature des positions qu'ils peuvent occuper dans la phrase. O. Bonami définit les positions syntaxiques suivantes au sein d'une « phrase matrice » [3] :

$$P_1 \quad sujet \quad P_2 \left\{ \begin{matrix} Verbe \\ Aux \quad P_3 \quad PPé \end{matrix} \right\} P_4 \quad objet \quad P_5$$

La position post-verbale (P_4) à droite du verbe est typiquement celle qui est occupée par les arguments, mais elle peut également accueillir des ajouts. La position post-verbe fini (P_3) est réservée aux ajouts [4], à l'exception de certains adverbes à valeur argumentale (*Il a bien traité ses invités*) et des arguments réalisés comme quantifieurs (*Il a tout dit*).

La position pré-V fini (P_2) peut être occupée par certains ajouts, mais pas par des arguments. La position en tête de phrase (P_1) peut être occupée par des ajouts : adverbes

2. Il n'est cependant pas sûr que cette version de la règle s'applique totalement, si l'on considère à la suite de Goldberg et Ackerman (2001) la notion d'ajouts obligatoires : *La maison a été construite* (*Ø | *l'année dernière*). Les paramètres de temps et d'aspect entrent également en jeu.

3. La plupart des exemples sont également empruntés à O. Bonami.

4. O. Bonami indique les acceptabilités suivantes :
Jean a, avec un marteau, cassé le vase : la présence d'un ajout est possible en position P_3.
**Jean a, à Marie, offert un disque* : la présence d'un argument est impossible dans cette même position.

de phrases, de cadre, et certains autres types d'ajouts. Mais elle ne leur est pas réservée : hormis les cas de dislocation gauche, des GP argumentaux peuvent également occuper cette position lorsqu'ils sont topicalisés :

(4) *Aux grévistes, la direction a dit qu'elle prendrait des mesures au plus vite.*

Au total, comme le résume O. Bonami :

> « Seule la position pré-V fini est réservée aux ajouts ; aucune position n'est réservée aux arguments. » (*op. cit.* , p. 57)

Test en *le faire* O. Bonami montre que *le faire* peut se combiner avec un ajout, mais pas avec un argument.

(5) a. *Jean est allé à Paris lundi soir, alors que Marie l'a fait mardi matin.*

b. ∗*Jean a rencontré son frère, alors que Marie l'a fait sa soeur.*

Au total, l'analyse de O. Bonami permet de conclure que la distinction doit être maintenue, puisqu'il existe des tests qui permettent de la valider de façon probante. Cette thèse tend à être corroborée sur le plan cognitif : dans un article de synthèse sur la question, (Tutunjian et Boland, 2008) font état d'expériences en psycholinguistique qui démontrent que, sur certains tâches de compréhension, les lecteurs consacrent moins de temps à la lecture des GP arguments qu'à celle des ajouts, et que cette différence de statut joue un rôle important dans le processus de désambiguisation. Elles précisent que ces éléments ne sont pas absolument incontestables, et signalent d'autres expériences qui n'étayent pas cette distinction. Elles concluent néanmoins positivement à la question posée dans le titre de leur article (« Do we need a distinction between arguments and adjuncts ? ») :

> « The evidence is not entirely conclusive on either front [formal linguistics and psycholinguistics], but on balance, the psycholinguistic evidence supports a formal distinction between arguments and adjuncts. » (*op. cit.*, p. 645)

6.2 Approche automatique

La question de la distinction entre arguments et ajouts a été abordée par tous les travaux qui visent l'acquisition de lexiques recensant les cadres de sous-catégorisation des verbes (*subcategorization lexicons*), comme Korhonen *et al.* (2006) pour l'anglais ou Messiant *et al.* (2008) pour le français :

> « A comprehensive subcategorization lexicon suitable for various NLP uses should firstly distinguish between arguments and adjuncts. »(Korhonen *et al.*, 2006)

Mais ces différents auteurs sont unanimes pour considérer que, mis à part des critères de fréquence, difficiles à contrôler, il n'est pas possible de disposer des informations nécessaires pour éliminer les ajouts des cadres de sous-catégorisation constitués à partir de corpus :

> « (...) currently many argument-adjunct tests cannot yet be exploited since they rest on semantic judgments that cannot yet be made automatically. » (*ibid.*)

La seule expérience qui vise précisément la distinction automatique entre arguments et ajouts est celle de Paola Merlo et ses collègues. Une première version de cette méthode a été présentée dans (Merlo et Leybold, 2001). Je présente brièvement la liste des tests linguistiques dont elles ont cherché à adapter le fonctionnement en corpus (sur le *Penn Treebank*).

dépendance vis-à-vis de la tête Cette mesure est estimée en calculant le nombre de verbes qui peuvent apparaître avec un GP donné. Leur corpus comportant trop peu de données, elles ont réalisé un étiquetage sémantique manuel des noms qui dépendent de la préposition, puis un étiquetage automatique à partir des classes de WordNet.

optionalité Cette mesure est calculée sur la base de la probabilité du GP d'apparaître avec le verbe ; une information relative à la classe des verbes est également intégrée dans le calcul.

itérativité Elle est estimée en calculant dans quelle proportion un GP trouvé en première position peut également être trouvé en seconde position lorsque le verbe a plusieurs dépendants prépositionnels.

La méthode est basée sur des procédures d'apprentissage. L'évaluation s'est faite sur la base des annotations disponibles dans le *Penn Treebank* : les GP avec un tag fonctionnel (comme *manner* ou *locative*) sont marqués comme ajouts, les autres comme arguments.

Les résultats montrent que la classification la plus efficace (86,5% de précision) intègre des informations relatives à la classe sémantique des mots. Cette méthode a été reprise dans (Merlo et Ferrer, 2006) et étendue aux dépendants du nom.

L'approche qui va être présentée ici présente des caractéristiques communes à celle de P. Merlo. L'idée de départ, tout d'abord, est la même :

> « the significance of this work resides in the exploration of the idea that the distinction between arguments and modifiers can be approached by capturing linguistic concepts statistically (...) The diagnostics (...) can be estimated by simple corpus counts. »(Merlo et Leybold, 2001)

Les entités sur lesquelles nous travaillons sont exactement de même nature :

> « the PP is approximated as the preposition and the PP-internal head noun. » (Merlo et Ferrer, 2006, 349)

P. Merlo et ses collègues vont cependant plus loin que nous ne l'avons fait dans la mesure où elles testent une plus grande variété de critères et s'appuient sur des annotations plus riches (en termes de classification sémantique en particulier) ; le fait de disposer de banques d'arbres syntaxiques annotées avec l'information qu'elles cherchent à détecter leur permet de mener une évaluation extensive de leur méthode.

De notre coté, comme nous allons le voir, nous nous sommes concentrés sur un critère principal (équivalente au critère dit de *head dependance*), et notre intérêt s'est déplacé du

repérage de l'opposition argument/circonstant vers l'examen de situations intermédiaires, du continuum.

6.3 Première approche : distinguer arguments et ajouts en corpus

Je me suis donc tout d'abord placée dans la perspective du repérage automatique de la distinction argument/ajout en corpus. Pour automatiser la différenciation entre les deux types de groupes prépositionnels, la première voie que nous avons envisagée a consisté à convertir en mesures applicables sur un corpus les tests linguistiques basés sur des manipulations syntaxiques ou sémantiques et des jugements d'acceptabilité.

6.3.1 Transposer les tests linguistiques

On l'a vu, dans leur grande majorité, les tests ne permettent pas de conclure de façon catégorique. Et ceux qui s'avèrent probants ne semblent guère opérationnels du point de vue de la linguistique de corpus, soit parce qu'on a peu de chances de trouver des occurrences des configurations à tester (test en *le faire*), soit parce que le critère s'applique sur des cas marginaux (le caractère obligatoire de certains arguments). Mais la possibilité de les utiliser pour des observations à grande échelle sur corpus change radicalement les conditions de leur applicabilité. On peut en effet espérer tirer parti de la masse de données pour dégager des tendances. O. Bonami lui-même, pour qui cette notion n'est pas opératoire, dit par exemple :

> « En français, les arguments (...) sont *typiquement* réalisés à droite du verbe qui les sélectionne » (*ibid.* p. 48, c'est moi qui souligne l'adverbe).

Il s'agit alors de déterminer quels tests peuvent être utilisés pour dégager des fonctionnements typiques, puis de sélectionner parmi eux ceux qui peuvent se prêter techniquement à une transposition en termes de tests applicables sur corpus.

Concernant le premier point, seul le test concernant le caractère obligatoire ou facultatif du complément ne se prête pas au jeu de la traduction en termes de typicité. Ce test est à manier avec beaucoup de précaution puisque, selon O. Bonami, les GP argumentaux ne sont pas typiquement obligatoires (seuls les arguments directs le sont). J'y ferai néanmoins appel dans l'expérience sur les adjectifs (section 6.5). Pour le reste (itérabilité, test en *le faire*, positions), il est assez aisé de dégager des tendances, et de dire par exemple : la capacité à occuper la place P_4 est typique de l'argument ; la capacité à occuper une grande variété de places est typique des ajouts.

Concernant la question de l'applicabilité concrète des tests, elle est plus coercitive, pour deux raisons de nature très différente. Le test peut faire appel à une manipulation qui produit une configuration qu'on n'a quasiment aucune chance de rencontrer dans un corpus. C'est le cas du test en *le faire*. Ou bien, le test fait appel à des informations que les sorties d'un analyseur syntaxique ne fournissent pas. C'est le cas du test d'itérabilité :

nous ne disposons pas de l'information qui permettrait de dire que c'est *le même type* (sémantique) de dépendant qui est reproduit.

Au total, la moisson est donc bien maigre. Seul le critère positionnel semble se prêter facilement à la transposition en vertu des deux critères que nous venons d'exposer. Nous l'exploitons partiellement dans une des versions de notre méthode (section 6.4.2). Mais au moment où nous lançons cette expérience avec C. Frérot nous sommes confrontées à un autre problème d'applicabilité, plus conjoncturel : dans cette phase du développement de SYNTEX, les positions préverbales ne sont pas encore suffisamment bien traitées. En particulier, l'analyse de la position P_3 repose sur une analyse fine du groupe verbal qui n'a été réalisée qu'ultérieurement.

Notre démarche a consisté alors à revenir aux définitions de la distinction entre argument et ajout pour tenter de forger d'autres tests, mieux adaptés aux outils d'exploration dont nous disposons dans une approche sur corpus.

6.3.2 Proposer de nouveaux tests

Les définitions que nous avons rappelées en section 6.1 caractérisent les ajouts par leur indépendance sur le plan syntaxique et sémantique par rapport au verbe, dans la mesure où ni leur position ni leur interprétation n'est conditionnée par lui. A l'inverse, les arguments sont contraints formellement et sémantiquement par le verbe. C'est donc une opposition entre **cohésion** et **autonomie** (syntaxique et sémantique) que nous cherchons à apprécier. Dans ce but, nous avons décidé de continuer à explorer les indications qu'offre le critère de productivité, tel qu'il avait été utilisé jusque là dans SYNTEX (cf. section 5.2.1). Ce critère permet en effet d'évaluer la régularité d'une relation : pour une configuration V GP donnée, une productivité élevée du couple V Prép est un indice de la régularité de cette association. Nous décidons d'utiliser à nouveau cet indice en considérant simultanément la productivité du couple Prép N, faisant ainsi l'hypothèse qu'elle permet de repérer l'existence d'une association régulière, qui n'est pas redevable de la relation à un verbe spécifique. Ce critère est à rapprocher de l'indice de dépendance à la tête utilisé par P. Merlo et ses collègues.

Nous avons esquissé dans (Bourigault et Fabre, 2000) et (Fabre et Bourigault, 2001) cette idée de tenir compte à la fois de la productivité du gouverneur et de celle du dépendant pour examiner le fonctionnement de la relation prépositionnelle. A partir des données issues de l'annotation d'un corpus de géomorphologie (275 000 mots), concernant à la fois des gouverneurs nominaux et verbaux, nous avons extrait les GP caractérisés par une productivité forte, c'est-à-dire associés à une grande variabilité de verbes. Tous les GP sont ramenés à un couple (p, n) qui en est le représentant, où p désigne la préposition et n le nom régi par la préposition. Ainsi, le couple $(sur, fond)$ est le représentant normalisé de tous les GP du corpus introduits par la préposition *sur* et dont la tête nominale est le nom *fond*. Le GP *sur le fond sablonneux* en est une instance spécifique, de même *sur des fonds qui varient de 5 à 10 mètres*. Ce représentant permet ainsi de rassembler des occurrences de différentes variantes d'un même GP. Notons que cette procédure peut cependant rapprocher accidentellement des instances que l'on souhaiterait distinguer. Ainsi, le couple $(à, coup)$ est-il instancié par des groupes ajouts comme *à coups de pioche* ou *à grands coups*, mais

aussi par le groupe adverbial *à coup sûr*.

La mesure de productivité du GP (p, n) est obtenue en calculant :

$$prod(p, n) = Card\{v / f(v, p, n) \geq 2\}$$

c'est-à-dire le nombre de verbes différents qui apparaissent dans des triplets (de fréquence au moins égale à 2) qui contiennent (p, n).

Dans cette première expérience, nous avons examiné les GP que SYNTEX a rattachés à un gouverneur verbal, et qui présentent la productivité la plus forte. Un premier examen de ces résultats m'a encouragée à vouloir systématiser cette approche, car la mesure met effectivement au jour des groupes autonomes, qui véhiculent pour l'essentiel une information de nature spatiale dans ce corpus (ex : *sur le fond*, *sur la pente*, avec une productivité de 15 dans les deux cas). Les observations que nous avons réalisées en appliquant la méthode au cas des gouverneurs nominaux sont de nature similaire, mais les résultats sont plus bruités, et les valeurs de productivité plus modestes. La méthode permet néanmoins de dégager quelques groupes productifs (ex : *à grain*, *à l'aval*, avec une productivité respectivement de 7 et de 3).

6.3.3 Première version de la méthode

C'est dans (Fabre et Frérot, 2002) qu'une méthode plus systématique, mais qui reste très préliminaire, est mise au point sur la base des observations précédentes. Elle consiste à nouveau à extraire du corpus l'ensemble des triplets (gouverneur, prép, dépendant) tels que le gouverneur est une forme verbale (infinitif, participe ou forme conjuguée) et le dépendant est la tête nominale du GP. On conserve également l'information relative à la présence ou l'absence du déterminant (cf. tableau 6.1), sachant le rôle, souvent décrit, du déterminant dans le fonctionnement des groupes prépositionnels (cf. l'opposition entre le programme de sous-classification en intension et en extension dont parle (Cadiot, 1997, p. 103 sq.) pour la préposition *à*). L'extraction est réalisée sur le même corpus de géomorphologie.

gouverneur	relation prépositionnelle	dépendant	déterminant
assimiler	*à*	*plan*	D
correspondre	*à*	*zone*	D
porter	*en*	*abscisse*	-
se produire	*sur*	*compartiment*	D
raisonner	*sur*	*échelle*	D
situer	*sur*	*plan*	D
varier	*le long de*	*ligne*	D

TABLE 6.1 – Triplets extraits par SYNTEX

Nous croisons ensuite les deux mesures de productivité en suivant l'algorithme suivant :

- Si, pour une préposition donnée, le gouverneur est productif ($prod(v,p) > 2$) et la productivité du dépendant est nulle ($prod(p,n) = 0$), alors le groupe prépositionnel du triplet (v,p,n) est étiqueté comme argument.
- Si, pour une préposition donnée, le dépendant est productif ($prod(p,n) > 2$ et la productivité du gouverneur ($prod(v,p) = 0$) est nulle, alors le groupe prépositionnel est étiqueté comme ajout.

1365 triplets sont ainsi étiquetés argument (ex : (*aboutir, à, aplanissement*)) et 525 triplets ajout (ex : (*allonger, dans, direction*)). Les résultats ont été évalués par deux méthodes parallèles. 50 triplets ont été sélectionnés aléatoirement parmi les GP étiquetés ajouts, 50 parmi les GP étiquetés arguments. Les 100 triplets ainsi retenus ont été triés par ordre alphabétique et soumis dans leur contexte d'origine (limité à une phrase) à une collègue linguiste [5], qui a caractérisé chaque GP comme argument ou circonstant. Nous avons également comparé les 100 triplets avec la description que donne le *Trésor de la Langue Française* des constructions syntaxiques dans lesquelles figure le gouverneur. Nous avons recherché la mention du GP dans l'article décrivant le verbe, celle-ci pouvant figurer dans différents champs descriptifs que nous avons définis (construction, crochets, exemple). Les résultats de cette évaluation sont présentés dans le tableau 6.2. On constate que 88 % des GP étiquetés argument par la méthode ont été identifiés comme tels par la juge, et 84 % d'entre eux sont décrits dans le TLFi. La précision est moins bonne pour les ajouts : 72 % des GP étiquetés ajouts ont été identifiés par la juge, 76 % étant absents des champs de description du TLFi que nous avons retenus.

décision	évaluation manuelle		comparaison avec le *TLF*	
	GP évalué argument	GP évalué ajout	GP décrit	GP non décrit
ARG	88%	12%	84%	16%
AJOUT	28%	72%	24%	76%

TABLE 6.2 – Evaluation

Les résultats montrent que la méthode offre un bon taux de précision. Mais elle ne permet de repérer que les cas bien tranchés, et le rappel est par conséquent extrêmement faible puisque la grande majorité des triplets combine des valeurs non nulles pour les deux mesures de productivité. En n'évaluant que les séquences qui présentent une valeur de productivité nulle, nous imposons une contrainte trop forte sur l'ensemble des triplets observables. Le reste des données, qui aurait pu être classé selon un gradient, n'a pas encore été examiné à ce stade.

5. Anne Le Draoulec (CLLE-ERSS).

6.4 Deuxième approche : calcul du degré d'autonomie du GP

Je suis revenue à cette question quelques années plus tard (Fabre et Bourigault, 2008). Nous disposions désormais de l'analyse de très grands corpus : les résultats sont issus du corpus LM10 (section 3.2.1). Avec l'aide de D. Bourigault, qui a beaucoup contribué à la mise au point des mesures que je vais présenter, je propose une méthode d'analyse plus complète.

6.4.1 Méthode

Pour calculer la force d'association entre le verbe et la préposition, nous faisons évoluer la mesure simple de productivité vers une mesure de productivité relative, notée $prod_R$, qui consiste à diviser la productivité du couple (v, p) par la productivité totale du verbe, notée $prod_T$(somme de ses productivités pour toutes les prépositions avec lesquelles il se construit). Cette mesure [6] permet d'atténuer les effets liés à la fréquence d'occurrence de la forme verbale et de mieux apprécier le rôle relatif de chaque préposition dans l'ensemble des GP rattachés au verbe .

Calcul de la productivité relative :

$$\begin{aligned} prod(v,p) &= Card\{n/f(v,p,n) \geq s, s = 2 \\ prod_T(v) &= \sum_p prod(v,p) \\ prod_R(v,p) &= \frac{prod(v,p)}{prod_T(v)} \end{aligned}$$

Prenons l'exemple de deux couples dont la productivité totale est identique :

$prod(insérer, dans) = 6 \qquad prod(songer, dans) = 6$

Leur valeur de productivité relative est par contre très différente, dans la mesure où le verbe *insérer* n'apparaît que dans 7 triplets ($prod_T(insérer) = 7$) (avec la contrainte $s \geq 2$), alors que le verbe *songer* apparaît dans 258 triplets, avec 7 prépositions différentes (majoritairement, la préposition *à*) ($prod_T(songer) = 258$). Cela donne les valeurs de productivité relative suivantes :

$prod_R(insérer, dans) = 0.86 \qquad prod_R(songer, dans) = 0.02$

De façon symétrique, nous nous intéressons à la propension du GP à se construire avec une diversité de verbes. Nous commençons par calculer une simple productivité du GP :

$prod(p, n) = Card\{v/f(v, p, n) \geq s, s = 2$

Dans un deuxième temps, nous prenons en compte la valeur de productivité relative précédemment calculée. L'idée est la suivante : il ne suffit pas de considérer le nombre de verbes avec lesquels le GP s'associe. En effet, ce nombre peut être élevé, mais si ces

6. Dans les articles, nous avons appelé cette mesure de productivité relative *selec*, car elle évalue le degré de sélection opéré par le verbe vis à vis de la préposition. Nous préférons ici nous en tenir à la dénomination $prod_R$, pour productivité relative.

verbes ont eux-mêmes tendance à sélectionner fortement la préposition ($prod_R$ haute en moyenne), on en conclut plutôt que le GP manifeste une forte cohésion avec ces verbes. C'est le cas pour le GP $(\grave{a}, question_D)$. Il se combine avec 27 verbes différents dans le corpus FRANTEXT20. La plupart présente une valeur de productivité relative forte avec la préposition *à* ;

$$\begin{aligned} prod_R(soustraire,) &= 1 \\ prod_R(consacrer,) &= 1 \\ prod_R(soumettre,) &= 0,99 \end{aligned}$$

On calcule donc une productivité pondérée par cette information :

$$prod_P(p,n) = \sum_{v/f(v,p,n)\geq 2}(1 - prod_R(v,n))$$

Au total, on aboutit à un calcul d'autonomie pour le GP :

$$auton(p,n) = \frac{prod_P(p,n)}{prod(p,n)}$$

Autrement dit : plus le rapport entre la productivité relative moyenne des verbes et la productivité du GP est bas, plus forte est l'indication qu'on a affaire à un GP autonome. Ainsi, le GP $(\grave{a}, dehors_D)$ a la même valeur de productivité que le GP $(\grave{a}, question_D)$ (27). Mais il s'associe à des verbes dont la productivité relative avec la préposition *à* est faible (ex : *tendre*, *manger*, *travailler*), alors que $(\grave{a}, question_D)$ s'associe à des verbes dont la valeur de $prod_R$ est haute (ex : *consacrer*, *soumettre*, *échapper*). L'autonomie du GP $(\grave{a}, dehors_D)$ est donc nettement plus élevée (0,8) que celle du GP $(\grave{a}, question_D)$ (0,3).

Nous différons au chapitre 7 la présentation des résultats de cette méthode de calcul, qui fournit une liste de GP ordonnée selon cette mesure d'autonomie. Nous disposons cette fois d'un moyen de calculer un degré d'autonomie, et non une opposition binaire, ce qui nous permet d'examiner non seulement les GP fortement cohésifs ou autonomes dans un corpus donné, mais également ceux qui se situent entre ces deux pôles, et permettent de questionner très empiriquement l'existence d'un continuum.

6.4.2 Introduction du critère de position

Dans (Fabre *et al.*, 2008), nous avons poursuivi ce travail en prenant également en compte le critère positionnel, grâce à la collaboration avec Lydia-Mai Ho-Dac, qui a travaillé sur la position initiale des GP. Nous avons cherché à voir s'il était possible d'opposer des GP statiques (apparaissant toujours en position postverbale), et des GP pouvant occuper des positions variées, et si ce critère de mobilité corroborait la mesure d'autonomie.

A partir des configurations données par O. Bonami (cf. section 6.1), nous avons décidé de privilégier les positions préverbales en tant que positions occupées typiquement par

des ajouts. Parmi elles, nous conservons seulement les positions P_1 (tête de phrase) et P_2 (pré-V fini), la position P_3 (post-V fini) étant difficile à prendre en compte dans la mesure où l'analyseur n'est pas toujours capable de repérer correctement les limites des complexes verbaux lorsque l'auxiliaire et le verbe sont séparés par un GP (*Il a, dans sa vie, toujours cherché à...*). On peut s'attendre par ailleurs à ce que peu de GP puissent être observés dans cette position.

Nous avons donc réalisé sur le corpus FRANTEXT20 les extractions des GP en position P_1 et P_2, soit 2948 couples en position P_1 et 1200 couples en position P_2, chaque couple comportant au moins 2 occurrences. Voici deux exemples :

(6) Position P1
GP $(dans, village_D)$: *Dans les villages que nous traversions, les paysans se tenaient...*

(7) Position P2
GP $(sur, gauche_D)$: *Un trou noir discret, sur la gauche, marque...*

Cette méthode a donc permis d'extraire à peine plus de 4000 couples, alors que le calcul d'autonomie livre, sur le même corpus, des résultats pour près de 21 000 couples. Les configurations étudiées concernent donc un nombre limité de types de GP ; elles ne permettent pas, par conséquent, de dire grand chose négativement des GP qui n'ont pas été trouvés dans ces positions – la non apparition d'une forme dans une structure globalement assez rare en corpus ne permet pas de conclure à son impossibilité.

Si l'on s'intéresse donc aux seuls GP qui passent les tests de mobilité – c'est-à-dire qui peuvent occuper des positions préverbales, on se rend compte qu'il se combine généralement avec une valeur d'autonomie haute, mais il est intéressant de remarquer que ce n'est néanmoins pas toujours le cas :

- l'autonomie moyenne des GP mobiles est de 0,76, contre 0,65 pour l'ensemble des GP post-verbaux ;
- un ensemble de GP se caractérisent à la fois par leur mobilité et une forte autonomie. C'est l'attendu. Les deux critères se confortent et permettent de repérer des GP qui sont clairement des ajouts. C'est le cas d'un GP comme $(dès, début_D)$: sa valeur d'autonomie est 0,99 et on le trouve aussi bien en position P_1 qu'en position P_2.
- Des GP peu autonomes présentent une forte mobilité. C'est par exemple le cas des couples $(à, étonnement_D)$ $(auton = 0{,}14)$ et $(à, parole_D)$ $(auton = 0{,}31)$. Chacun d'eux apparaît bien en position préverbale :

(8) *Mais le récit, à son propre étonnement, avait tourné d'une autre manière.*

(9) *Aux paroles sèches de Naroumof, elle se leva, se dirigea vers Anne.*

Cette information contradictoire s'explique par le fait que, lorsque ces couples s'instancient en position postverbale, ils entrent cette fois dans la portée du verbe (*il se mêlait à mon étonnement* ; *il avait peine à s'intéresser aux paroles...*). Le décalage entre les deux indices s'explique donc par le fait qu'on n'a pas affaire aux mêmes types d'instances avant et après le verbe. Contrairement au cas du GP *dès le début*, qui s'émancipe systématiquement du verbe de la proposition quelle que soit sa position, les GP représentés par les couples $(à, étonnement_D)$, $(à, parole_D)$ n'ont un statut de complément de phrase qu'en position

préverbale dans le corpus. Le retour aux instances particulières s'avère donc indispensable pour affiner la description obtenue sur des entités parfois trop générales.

6.5 Extension à l'étude de la complémentation adjectivale

Je présente ici les résultats de la collaboration avec Anna Kupść (Fabre et Kupść, 2009), qui m'a permis d'explorer les possibilités d'appliquer au cas des compléments de l'adjectif une méthode semblable, basée sur l'utilisation d'indices de caractérisation du fonctionnement des GP à partir de grands corpus annotés par SYNTEX. L'objectif d'A. Kupść est de constituer un lexique des adjectifs en français comportant des informations de sous-catégorisation. Nous avons ainsi pu combiner deux approches des corpus pour assister cette tâche.

6.5.1 Spécificités de la complémentation adjectivale

La tâche d'identification des compléments des adjectifs se trouve encore compliquée par rapport au verbe par plusieurs spécificités de la sous-catégorisation adjectivale. Tout d'abord, celle-ci entre en concurrence avec des constructions syntaxiques englobantes (comme dans l'exemple : *il est trop fier pour parler*, où *pour* dépend de la structure *trop* X *pour* VINF). Il faut donc différencier les dépendants des adjectifs et les GP qui dépendent de gouverneurs situés plus haut dans l'arbre syntaxique. Concernant la distinction entre arguments et ajouts, elle se heurte en outre à deux particularités : les compléments des adjectifs sont généralement omis. Le caractère optionnel ou obligatoire du dépendant est donc encore moins opérant ; la variabilité de la préposition est très importante, y compris dans le cas des arguments (par exemple, le complément de l'adjectif *aimable* peut être introduit indifféremment par *envers* ou par *à l'égard de*).

6.5.2 Combiner deux approches sur corpus

6.5.2.1 Approche par *treebank*

Le *French Treebank* (désormais FTB), est un corpus arboré de 1 million de mots, développé au LLF (Université Paris 7), qui comprend 4 années du journal *Le Monde*. L'annotation linguistique a été réalisée sur la base d'un traitement automatique initial, qui a ensuite été validé manuellement. Elle donne accès à une segmentation en mots et en constituants syntaxiques, à une catégorisation morpho-syntaxique des mots et groupes de mots (lemmes et propriétés morphologiques), et à une caractérisation des fonctions syntaxiques de certains groupes de mots dans la phrase.

Une approche à base de règles linguistiques (Kupść, 2008) consiste à extraire les *frames* (cadres de sous-catégorisation) des adjectifs. Les informations sont nettement moins riches que dans le cas des verbes (Kupść, 2007) puisqu'on ne dispose pas d'annotation fonctionnelle. Les dépendants de toute nature sont donc récupérés. En outre, malgré le fait que le FTB a été annoté manuellement et validé, il reste inévitablement des erreurs de rattache-

ment. En aval de l'extraction des dépendants, il s'agit donc de repérer, pour les éliminer, deux types de cas :

- le dépendant est en fait contrôlé par une structure syntaxique englobante et ne relève pas de la valence adjectivale (ex : *le plus beau du monde*),
- le dépendant est un ajout (ex : *sinistré après la tempête*).

Diverses heuristiques sont utilisées pour identifier les constructions qui peuvent interférer avec la valence adjectivale, en particulier les constructions comparatives (*La réunion était plus intéressante que je ne pensais*), superlatives (*le plus anglais des Irlandais*) ou intensives (*C'est trop beau pour être vrai*). Par ailleurs, dans la mesure où le corpus ne fournit pas d'élément pour guider la distinction entre arguments et ajouts, (Kupść, 2008) opte pour une élimination des prépositions qui ne sont pas susceptibles d'introduire des arguments[7].

Le traitement fournit 40 *frames* pour 2153 adjectifs. Parmi eux, 1849 adjectifs apparaissent uniquement avec le *frame* élémentaire SUJ:NP (aucun dépendant de l'adjectif). Seuls 304 adjectifs entrent dans des cadres de sous-catégorisation plus élaborés, comme par exemple :

reconnaissant SUJ:NP, OBJ:PP[envers]
fier SUJ:NP, OBJ:PP[de]

6.5.2.2 Limites de l'approche par *treebank*

L'analyse des structures extraites montre deux pistes d'amélioration de l'extraction :

1. on manque d'indices pour repérer correctement les constructions syntaxiques concurrentes à la dépendance adjectivale ;
2. on manque d'éléments probants pour distinguer les arguments des ajouts.

Prenons le cas des impersonnels pour illustrer le premier point. Dans le cas d'une structure impersonnelle (ex : *Il est absurde de ne pas le reconnaître*), il s'agit de transformer l'information extraite, dans laquelle l'infinitive est identifiée à tort comme un dépendant de l'adjectif :

absurde SUJ:NP, OBJ:VINF[de]

en un cadre dans lequel l'infinitive est cette fois correctement identifiée comme le complément sujet de l'adjectif :

absurde SUJ:VINF

Le recours à un grand corpus est alors envisagé comme un moyen de filtrer les candidats extraits et d'augmenter le nombre d'adjectifs décrits, voire le nombre de *frames*. Cette étape pourrait certainement être réalisée entièrement manuellement. Cette démarche d'acquisition automatique d'indices pour assister le travail de filtrage vise à apprécier la validité des mesures que nous voulons utiliser dans un deuxième temps pour extraires de nouveaux *frames* candidats du corpus LM10.

7. Elles ne figurent pas dans le lexique PREPLEX (Fort et Guillaume, 2007) utilisé par A. Kupść pour disposer d'une liste de prépositions argumentales.

6.5.3 Filtrer les patrons candidats

Nous nous sommes concentrées dans cette phase sur l'extraction des GP, en incluant les infinitifs prépositionnels, mais en excluant les compléments propositionnels, pour deux raisons : l'extraction à partir du FTB s'avère la plus problématique dans le cas des GP, et ce sont les types de compléments adjectivaux que l'on a le plus de chances de récupérer de manière fiable en sortie de SYNTEX. Je retrace les étapes de la méthode consistant à détecter, sur les sorties du corpus LM10 analysé, des éléments permettant de mieux caractériser les patrons candidats issus du FTB.

6.5.3.1 Extraction des triplets

Nous extrayons les dépendants des adjectifs identifiés par SYNTEX lorsque ceux-ci sont des GP. Nous retenons deux types de patrons :

1. ADJ P N
2. ADJ P VINF

Cette extraction est effectuée pour tous les adjectifs, après réduction des coordinations éventuelles et quelle que soit la relation dont ils sont la cible (tableau 6.3).

adj	prép	dép	cat-dep	relation	suj	seq
impatient	de	assister	VINF	ATTS	Je	Je suis impatient d'assister ...
idéal	pour	cyclo-cross	N	ATTS	Ce	C'est idéal pour le cyclo-cross...
interne	à	parti	N	ADJ	document	Dans un document interne au parti...

TABLE 6.3 – Patrons de complémentation adjectivale

Nous utilisons des heuristiques pour filtrer les constructions qui peuvent l'être : constructions superlatives (*le plus anglais des Irlandais*) ou marquant l'intensité (*assez souple pour supporter*).

Le gain en volume de données est manifeste par rapport à l'extraction issue du FTB, comme on le voit dans le tableau 6.4.

	Treebank	LM10
nb d'adjectifs (types)	304	2684
patrons Prep (VINF|N)	26	136
patrons Adj Prep (VINF|N)	369	6778

TABLE 6.4 – Comparaison *French Treebank* / corpus annoté automatiquement

Mais ce volume s'accompagne évidemment de beaucoup de bruit, dû aux erreurs de l'analyse automatique, qu'il s'agisse d'erreurs d'étiquetage (*sauf* est étiqueté adjectif dans

sauf par grand froid), de segmentation (*au contraire* n'est pas identifié comme adverbe, ce qui déclenche l'acquisition du triplet *polémique à N* à partir de la séquence *polémique au contraire*) ou d'analyse syntaxique (le triplet *ponctuel à N* est calculé à cause d'un mauvais rattachement de la préposition *à* dans *offrir une aide ponctuelle à des élèves*). Il s'agit donc de fournir des éléments d'information complémentaires pour extraire de ces données des informations utiles à la constitution d'un lexique syntaxique.

6.5.3.2 Identifier les structures impersonnelles

Le volume de données permet parfois d'employer des procédures qui allient une très grande simplicité et une très grande efficacité. C'est le cas du repérage de structures impersonnelles. Pas plus que dans le FTB nous ne sommes en mesure de déterminer localement, sur la base de l'analyse fournie par SYNTEX, si l'on a affaire à un sujet personnel ou impersonnel lorsque l'adjectif est attribut du sujet. Les analyses produites sont les mêmes dans l'exemple 10 et 11 : le pronom reçoit la même étiquette PRO, il est sujet du verbe *être*, qui gouverne l'adjectif selon la relation ATTS, et le groupe infinitif est rattaché pareillement à l'adjectif par la relation PREP.

(10) `<TXT>`Il est également assez déconcertant de constater que ...
`<ETIQ> Pro|il|Il|1|SUJ;2| VCONJS|être|est|2||SUJ;1,ATTS;5`
`Adv|également|également|3|ADV;5| Adv|assez|assez|4|ADV;5|`
`Ppr|déconcerter|déconcertant|5|ATTS;2|ADV;3,ADV;4,PREP;6`
`Prep|de|de|6|PREP;5|NOMPREP;7 VINF|constater|constater|7|NOMPREP;6|OBJ;8`

(11) `<TXT>`Il est maintenant capable de tout construire ou de tout détruire
`<ETIQ> Pro|il|Il|1|SUJ;2| VCONJS|être|est|2||SUJ;1,ATTS;4`
`Adv|maintenant|maintenant|3|ADV;4| Adj??|capable|capable|4|ATTS;2|ADV;3,PREP;8`
`Prep|de|de|5|CC;8|NOMPREP;7 Pro|tout|tout|6|| VINF|construire|construire|7|NOMPREP;5|`

Mais la masse de données constitue dans ce cas un atout qu'il est facile d'exploiter : pour chacun des 40 *frames* de type VPinf[*de*] extraits du FTB, on calcule un indice *impers*, qui fournit la proportion de sujets susceptibles d'avoir une valeur d'impersonnel – formes *il*, *ce*, *cela*. Malgré l'ambiguïté du pronom *il*, le résultat permet de conclure dans quasiment tous les cas :

- 27 *frames* candidats ont une valeur très haute de l'indice : *impers* > 0,9
  Ex : *Il est* (*absurde*, *acceptable*, *anormal*) *de* + VPinf
- 12 *frames* candidats ont une valeur très basse : *impers* < 0,15
  Ex : *Il est* (*capable*, *conscient*, *content*) *de* + VPinf

Seul le patron *nouveau*[$_{\text{VINF}}$ de] reçoit une valeur intermédiaire (*impers* = 0,4). Bien qu'il s'agisse d'une structure impersonnelle dans tous les cas, son repérage est compliqué par la présence de structures spécifiques, dans lesquelles un élément nominal se substitue au pronom personnel :

(12) a. *L'idée n'est pas nouvelle de mettre...*

b. *L'utopie n'est pas nouvelle de pimenter ...*

La méthode d'identification des constructions impersonnelles, sur les *frames* candidats, a donc une excellente performante (100% de précision, 93% de rappel), parce que la décision à prendre est binaire pour chaque patron, et qu'elle peut être prise à partir d'indices très simples qui s'avère fiables dès qu'on dispose d'un nombre de données suffisant.

6.5.3.3 Identifier des structures douteuses

Notre projet était d'adapter aux adjectifs la méthode mise en œuvre pour les verbes, de manière à fournir des éléments permettant de distinguer des GP cohésifs et autonomes, et de repérer ainsi les groupes susceptibles de ne pas relever de la valence de l'adjectif.

Nous avons d'abord calculé les mesures de productivité et de productivité relative sur le corpus LM10 des couples (*adj,p*) extraits du FTB. Le complémentaire – calcul du degré d'autonomie du GP – est apparu peu adapté à l'étude des adjectifs. Alors que dans le cas des ajouts verbaux, on sait qu'on a affaire à des groupes qui peuvent être mobiles et s'adjoindre plus haut dans la structure syntaxique (et en particulier au niveau de la phrase), il est difficile de concevoir ce que seraient des GP autonomes vis à vis de l'adjectif.

Nous avons également reformulé le critère d'optionalité en calculant dans quelle proportion un adjectif est trouvé avec un dépendant dans le corpus : *presence_*GP(adj). Le résultat confirme la spécificité de la valence adjectivale : seule une petite minorité (3,7 %) des adjectifs du corpus sont trouvés par l'analyseur avec un dépendant dans plus de 50 % des cas (ex : *désireux*, *inhérent*, *prompt*) .

Munis de ces deux mesures (productivité et optionalité), nous avons extrait des *frames* candidats ceux qui présentent des valeurs basses : le patron est peu productif ($prod<10$), l'adjectif est rarement trouvé avec une expansion prépositionnelle ($presence\_\text{GP}<0{,}1$), et lorsqu'il l'est l'expansion n'est pas souvent introduite par cette préposition ($prod_R<0{,}1$). Ce filtre met au jour 24 *frames* candidats. Leur examen montre qu'il ne s'agit en effet pas de patrons valides mais :

- de structures mal formées, faute en particulier d'un repérage d'une structure syntaxique englobante (*étroit*[VINF pour], *efficace*[VINF à])
- d'ajouts (*dynamique*[PP depuis], *courant*[EN de])

Le seuil employé est forcément arbitraire. Mais cette première phase nous a montré que ce type de mesures (indice *impers*, mesures *prod*, $prod_R$, *presence_*GP) permettait de repérer des intrus au sein d'une liste de patrons acquis à l'aide d'une procédure d'extraction pourtant très précisément controlée (corpus source validé manuellement, règles de filtrage en aval). Nous avons donc prolongé ce travail par une deuxième phase consistant cette fois à utiliser le corpus LM10 comme source pour l'extraction de nouveaux patrons.

6.5.4 Extraire de nouveaux patrons

6.5.4.1 Eliminer les structures impersonnelles

Etendue aux 431 patrons PREP VPINF extraits de LM10, la méthode calcule une valeur *impers* très haute (>0,8) pour 84 % des patrons, très basse (<0,2) pour 7% d'entre eux. L'examen de ces résultats montre que dans tous ces cas, l'indice permet de discriminer correctement les deux constructions. Il a donc un rappel de 91 % et une précision de 100 %.

9 % des patrons ont une valeur intermédiaire et demandent un examen plus approfondi. On retrouve le cas de *nouveau*, c'est-à-dire des impersonnels qui apparaissent dans des structures alternatives :

(13) *c'est lourd de lui faire porter...*

(14) *l'erreur serait lourde de le cantonner...*

On trouve également des cas de structures ambiguës :

(15) *Julie est trop malheureuse d'avoir été* (construction personnelle)

(16) *c'est malheureux de parler de ça* (construction impersonnelle)

Il faut noter que certains adjectifs n'apparaissent jamais dans une construction impersonnelle dans le corpus (c'est par exemple le cas d'*aimable*) alors qu'ils sont compatibles avec cette construction. On retrouve ici les limites bien connues des corpus, même grands, pour l'acquisition de ressources lexicales génériques.

6.5.4.2 Acquérir de nouveaux patrons

Le tableau 6.5 donne une idée du décalage entre les données brutes extraites du corpus LM10 (nous extrayons tous les patrons candidats, de fréquence supérieure à 2), et les *frames* extraits du FTB.

	frames candidats FTB	patrons LM10
productivité	293	20
$prod_R$	0,5	0,3
*presence_*GP	25%	4%

TABLE 6.5 – Comparaison des valeurs moyennes obtenues sur TB et LM10

Notre démarche a consisté à extraire de cette masse de données les patrons présentant les propriétés optimales pour être de bons candidats, à savoir :

- le patron est productif
- l'adjectif est souvent trouvé avec une expansion prépositionnelle
- l'expansion prépositionnelle que l'on trouve pour l'adjectif est majoritairement introduite par la préposition considérée

En assignant des valeurs très hautes aux mesures correspondant à ces trois critères (*prod*>5, $prod_R$>0,2, *presence_*GP>20%) (hautes au vu des valeurs moyennes données dans le tableau 6.5), nous avons extrait 199 patrons, qu'un examen manuel permet d'analyser de la façon suivante :

- 51 % sont des arguments.
  Ex : *conforme* PP[*à*] (valeurs : 544/0,97/0,68)
- 17 % ne correspondent pas à une information de valence, même s'il s'agit d'associations très régulières (*valable jusqu'à*, *consultable sur*, *payable par*).
- 32 % relèvent d'erreurs d'étiquetage et d'analyse.

L'extraction est donc bruitée, mais une sélection manuelle permet de récupérer ainsi 101 nouveaux *frames*, augmentant la ressource initiale issue du FTB de 27%.

Beaucoup d'autres patrons sont intéressants. Ce sont en particulier ceux qui figurent dans les 17 % d'associations non argumentales mais régulières. Pour beaucoup d'entre eux il est difficile de se prononcer quant à leur statut. Les mesures que nous avons calculées fournissent des indications quant à leur fonctionnement dans le corpus, des indices de régularité. Il est par exemple intéressant de repérer des patrons qui présentent des valeurs hautes alors que la préposition n'est pas considérée susceptible d'introduire des arguments (si l'on considère la ressource PREPLEX évoquée plus haut). Nous en donnons des exemples dans le tableau 6.6.

ADJ	PRÉP	$prod_R$	*prod*	*presence_*GP
indisponible	*pendant*	0,22	7	0,26
livrable	*jusqu'à*	0,21	17	0,49
méfiant	*à l'égard de*	0,6	103	0,2
sceptique	*quant à*	0,24	25	0,26

TABLE 6.6 – Patrons non argumentaux mais présentant des indices forts de régularité

Le statut de ces données est indissociable d'une réflexion sur l'utilisation de la ressource qui est construite : le fait de s'en tenir à une distinction binaire entre ce qui est argumental et ce qui ne l'est pas peut priver un système de TAL d'informations utiles, qui amènent à se représenter le fonctionnement des compléments de manière plus graduée. Le point de départ de l'étude consistait à n'intégrer que les compléments sous-catégorisés. Le fait de capter des indices en corpus nous a amené à nous confronter une fois de plus au problème du continuum.

6.6 Conclusion

L'objectif premier de ce travail était de se doter de moyens d'apprécier de façon automatique la distinction argument-circonstant, et de faciliter les décisions de désambiguisation locale réalisées dans l'analyseur. L'intérêt et l'originalité de l'approche tiennent au fait que l'on peut désormais tenir compte de la nature du GP et s'appuyer sur la mesure de son autonomie pour orienter les décisions d'attachement prépositionnel. A titre d'exemple, le

rattachement d'un GP en *à* pour le verbe *demander* peut dépendre de la valeur d'autonomie calculée pour ce GP :

(17) a. *demander une signature à l'arrivée* : *auton* ($à$,$arrivée_D$)=0,68
b. *demander une signature au client* : *auton* ($à$,$client_D$)=0,38

Autre exemple pour le verbe *viser* :

(18) a. *viser à la tête* : *auton* ($à$,$tête_D$)=0,646
b. *viser au maintien* : *auton* ($à$,$maintien_D$)=0,265

L'expérience ne s'est pas conclue par une implémentation dans SYNTEX qui permettrait d'évaluer cette contribution sur les performances de l'analyseur. De fait, mon intérêt s'est déplacé vers l'utilisation de ces éléments de quantification pour analyser les GP d'un corpus, et explorer les cas d'autonomie médiane, puisque cette mesure fournit un gradient permettant d'explorer le continuum entre arguments et circonstants. Le chapitre suivant présente les résultats obtenus dans la perspective d'une analyse linguistique de la complémentation.

Chapitre 7

Exploration quantifiée de la complémentation verbale

Je consacre ce chapitre à l'analyse des résultats obtenus par les méthodes d'exploration que je viens de décrire, de manière à montrer comment ils peuvent alimenter une analyse descriptive des constructions verbales basée sur les corpus. Je suggère les pistes d'observation qui se dégagent lorsque l'on examine de façon plus détaillée les résultats obtenus par cette approche inductive de la complémentation verbale, fondée sur l'exploration systématique de grandes collections de textes et le recours à des mesures statistiques qui permettent d'aborder sous l'angle de la quantification certaines propriétés des compléments verbaux habituellement évaluées à l'aide de tests linguistiques de grammaticalité.

Cette étude, menée en collaboration avec Josette Rebeyrolle [1], consiste à travailler à la caractérisation de la complémentation verbale en se fondant sur l'analyse du comportement des verbes et de leurs compléments prépositionnels dans le corpus. On bénéficie ainsi d'un dispositif d'observation qui permet de dégager des profils d'usage : profils de verbes (section 7.1) ou de prépositions (section 7.3), profils de GP selon leur degré d'autonomie (section 7.2). Par ailleurs, comme nous avons fait le choix de travailler sur deux corpus différents (LM10 et FRANTEXT20), nous sommes en mesure d'apprécier la part de stabilité et de diversité dans les comportements des compléments selon les types de textes (section 7.4).

Plus précisément, je considère la façon dont les mesures de productivité relative et d'autonomie permettent d'explorer les différents objets étudiés et d'en dresser des profils :

- profils des verbes : la mesure de productivité relative fournit des éléments sur la relation entre verbes et prépositions ;
- profils des groupes prépositionnels : ils s'organisent entre GP autonomes et cohésifs ;
- profil des prépositions : on peut différencier les emplois des prépositions selon la tendance qu'ils montrent à introduire des GP plus ou moins autonomes.

1. Il a fait l'objet d'une communication au colloque de l'AFLICO, au sein de l'atelier « La linguistique fondée sur l'usage », animé par Jacques François et Dominique Le Gallois. Un projet d'article a été accepté pour parution dans le numéro thématique de la revue *Travaux linguistiques* intitulé « Réflexions et arguments pour une linguistique fondée sur / guidée par l'usage », et coordonné par ces mêmes collègues.

7.1 Profils de verbes

La mesure de productivité relative ($prod_R$) fournit les premiers éléments d'observation. Elle permet de dégager différents profils de verbes vis-à-vis des prépositions qui leur sont rattachées par l'analyseur.

Le premier type de verbes entretient une relation exclusive avec une préposition particulière. La valeur de $prod_R$ est 1, ou proche de 1 (cf. tableau 7.1 – valeurs issues de FRANTEXT20). Dans ce cas, on a affaire, en général, à une relation argumentale claire, comme pour *comparer à, s'attacher à* ou *hériter de*. Dans d'autres cas, la relation peut ne pas être classiquement décrite comme argumentale ; c'est le cas de *patauger dans* ou *jongler avec*. Mais la mesure indique que l'expression d'un complément spatial ou d'un complément de manière est très étroitement associée au verbe. Elle signale ainsi le caractère obligatoire du complément, sinon sur le plan syntaxique (il faudrait étudier les possibilités d'effacement du GP), du moins sur le plan sémantique (cf. 6.1).

verbe	préposition	$prod_R$
comparer	*à*	1
s'attacher	*à*	1
hériter	*de*	1
patauger	*dans*	1
jongler	*avec*	1

TABLE 7.1 – Profil 1 : relation exclusive avec une préposition

Le deuxième type de verbes est associé à plusieurs prépositions, dont l'une domine (cf. tableau 7.2 – valeurs issues de LM10). Ce profil correspond à plusieurs cas de figure : la relation prédominante (ex : *témoigner de* <quelque chose>) s'accompagne de dépendants plus périphériques qui ajoutent d'autres informations au verbe (*devant* <une assemblée>) ; ou bien une même relation peut être exprimée de façon majoritaire à l'aide d'une préposition (*causer avec*, *accrocher à*), et plus marginalement avec une autre (*causer à*, ou *accrocher dans*). On trouve aussi des cas de polysémie (*se mêler de*, *se mêler à*).

verbe	préposition	$prod_R$	autres prépositions
témoigner	*de*	0,84	*devant*
accrocher	*à*	0,69	*dans*
se mêler	*à*	0,67	*de*
causer	*avec*	0,49	*de, à*
arriver	*à*	0,43	*dans, en*

TABLE 7.2 – Profil 2 : plusieurs prépositions, l'une domine

Le troisième type de verbes (tableau 7.3) présente une concurrence plus marquée entre plusieurs prépositions.

verbe	préposition 1	prod_R(verbe,prép1)	préposition 2	prod_R(verbe,prép2)
s'accorder	*à*	0,45	*avec*	0,45
ruisseler	*de*	0,33	*sur*	0,39
hésiter	*entre*	0,38	*sur*	0,35
se tromper	*de*	0,45	*sur*	0,36

TABLE 7.3 – Profil 3 : concurrence entre plusieurs prépositions

Là encore, ce cas de figure renvoie à des fonctionnements divers. On repère des cas de constructions alternatives dont l'usage semble être équilibré dans le corpus avec les verbes *s'accorder* (*à* / *avec*), *hésiter* (*entre* / *sur*), *ruisseler* (X *ruisselle de* Y / Y *ruisselle sur* X). On a également affaire à des emplois sémantiquement distincts (*se tromper de* / *se tromper sur*).

Cette mesure permet donc de repérer un premier niveau de fonctionnement, limité au degré d'association entre le verbe et la préposition.

7.2 Profils de groupes prépositionnels

La méthode que nous avons mise en œuvre [2] fournit également une liste de GP ordonnée selon la mesure d'autonomie que nous avons calculée. C'est une manière plus inhabituelle de considérer nos résultats : si le degré d'association entre un verbe et une préposition est une donnée que l'on a l'habitude d'appréhender, il est nouveau de se représenter le GP selon une tendance à la cohésion ou à l'autonomie sur l'ensemble d'un corpus donné. Nous allons voir que la mesure d'autonomie permet tout d'abord de dégager des GP situés aux deux extrémités du spectre, autrement dit des GP fortement autonomes ou fortement cohésifs. Nous nous intéresserons ensuite à des situations médianes.

7.2.1 GP autonomes, GP cohésifs

SYNTEX dispose, en aval de l'étiquetage réalisé par TREETAGGER, d'une procédure de segmentation et d'étiquetage visant les groupes prépositionnels dont la fonction est assimilée à celle d'un adverbe dans la phrase. Cela permet d'éviter de chercher à analyser ces unités au sein du module de rattachement prépositionnel. A titre d'exemple, sont étiquetées comme ADVGP les séquences suivantes : *par ricochets*, *sans doute*, *au demeurant*, *de fait*, *en effet*, etc. Notons que cette technique n'est pas idéale, car cette décision est très souvent sensible au contexte. Ainsi, la séquence *à la baisse* est étiquetée ADVGP, ce qui se justifie dans l'exemple suivant :

2. Nous analysons ici les résultats obtenus en appliquant le critère d'autonomie, sans utiliser le travail sur les positions que nous avons décrit au chapitre précédent.

(1) *Tout le plan de restructuration d' Hachette consistait à revoir à la baisse les ambitions de la chaîne (...)*

mais est inapproprié dans un grand nombre de cas, car la préposition peut être sous-catégorisée, comme dans l'exemple suivant :

(2) *(...) certains commentateurs considèrent que les maux dont souffre l'économie américaine exigent des remèdes à très long terme seuls susceptibles (...) de* ***parer à*** *la baisse de la productivité.*

L'information relative au degré d'autonomie des GP dans le corpus peut donc fournir un indice utile à la segmentation et à l'analyse syntaxique automatiques. De fait, si l'on s'intéresse aux GP qui ont une valeur *auton* très haute, on constate qu'ils requièrent en effet un traitement particulier, et correspondent à deux types d'unités[3] :

1. Des **locutions prépositionnelles** : ainsi, parmi les GP les plus autonomes associés à la préposition *à* figurent les séquences *à l'égard* (*de*), *à l'échelle* (*de*), *au profit* (*de*), *au matin* (*de*), *à parité* (*de*). Leur valeur d'autonomie se situe autour de 0,8.
2. Des **locutions adverbiales** : par exemple, parmi les GP les plus autonomes associés à la préposition *de* figurent les séquences *en majorité*, *en contrepartie*, *en finale*, *en politique*, *en tête*, *en réseau*, etc. Leur valeur d'autonomie est supérieure à 0,9.

Certains des GP très autonomes peuvent relever simultanément des deux cas de figure (ex : *en cours* (*de*), *en tête* (*de*)).

La mesure d'autonomie peut donc faciliter le recensement de ces unités.

Outre ce cas de figure, on a plus généralement affaire, au niveau du pôle d'autonomie positif, à des séquences syntaxiquement et sémantiquement autonomes, à la sémantique stable, selon les termes de Grimshaw (1992). Le tableau 7.4 (tiré de (Fabre *et al.*, 2008)) donne quelques illustrations pour les prépositions *à*, *dans*, *de* tirées du corpus FRANTEXT20.

à	*auton*	*de*	*auton*	*dans*	*auton*
($à$,$horizon_D$)	0,79	(de,ton_D)	0,87	($dans$,$style_D$)	0,91
($à$,$vitesse_D$)	0,77	(de,$trait_D$)	0,82	($dans$,$langue_D$)	0,89
($à$,$lueur_D$)	0,75	(de,$manière_D$)	0,82	($dans$,an_D)	0,86
($à$,$crayon_D$)	0,75	(de,$voix_D$)	0,81	($dans$,$moment_D$)	0,85
($à$,$surface_D$)	0,79	(de,$façon_0$)	0,78	($dans$,$jeunesse_D$)	0,85

TABLE 7.4 – Exemples de GP autonomes avec les prépositions *à*, *de*, *dans*

On constate que ces groupes sont porteurs d'informations circonstancielles caractéristiques du corpus traité, relatives ici au temps (*dans sa jeunesse*), à la manière (*d'un ton* (+ adj)), à l'espace (*à l'horizon*).

A l'autre extrémité du continuum, on relève à l'inverse des groupes très cohésifs, illustrés pour les prépositions *à* et *de* dans le tableau 7.5.

3. Pour simplifier la lecture, je remplace parfois la notation générique (PREP, NOM) par une instance particulière (la plus fréquente) du modèle de GP. Par exemple, je note *à droite* pour signifier ($à$,$droite_0$), sachant que d'autres instances sont possibles, comme *à droite de l'écran*.

à	*auton*	*de*	*auton*
(*à*,$effort_D$)	0,12	(*de*,$feuille_0$)	0,09
(*à*,$volonté_D$)	0,14	(*de*,$nuage_D$)	0,14
(*à*,$exigence_D$)	0,15	(*de*,$voile_D$)	0,18
(*à*,$émotion_D$)	0,18	(*de*,$droit_D$)	0,18
(*à*,$influence_D$)	0,19	(*de*,$spectacle_0$)	0,19

TABLE 7.5 – Exemples de GP cohésifs avec les prépositions *à* et *de*

On voit que ces GP sont fortement dépendants du verbe dans la mesure où il est difficile d'évaluer leur apport sémantique propre. Le rétablissement d'un verbe particulier est nécessaire à l'interprétation (*renoncer à l'effort*, *tapisser de feuilles*, *se perdre dans la contemplation*).

7.2.2 Etude du continuum

A partir de (Fabre et Bourigault, 2008), je me suis intéressée à l'examen de la zone intermédiaire, c'est-à-dire à des GP dont le taux d'autonomie est médian. Cette zone médiane est assimilée aux valeurs $0,3 \rightarrow 0,7$ de la mesure *auton*. Les GP rencontrés dans cette zone ont principalement deux comportements possibles.

7.2.2.1 Profils contrastés

Il peut tout d'abord s'agir de GP qui présentent un profil contrasté, jouant dans le corpus tantôt le rôle d'arguments, tantôt le rôle d'ajouts. On trouve en particulier ce cas de figure avec la préposition *à* et la préposition *dans*. Un GP comme ($à, portière_D$) s'associe à la fois à des verbes qui sous-catégorisent fortement la préposition (*s'adosser*, *s'accouder*) et à des verbes pour lesquels le complément en *à* a le rôle d'ajout (*regarder*, *apparaître*). De même pour un GP comme ($dans, sommeil_D$), on trouve à la fois des verbes dont le taux de sélection est très fort pour *dans (sombrer, plonger, s'enfoncer*), et des verbes pour lesquels l'information spatiale est périphérique (*goûter, se retourner, abandonner,* etc.) – dans ce cas, le GP prend une valeur temporelle.

Ces cas médians relèvent d'une rencontre entre des types de complémentation distincts et relativement bien tranchés, pour lesquels la prise en compte du verbe déplace l'analyse vers un pôle ou vers l'autre. Ils ne sont pas les plus pertinents pour travailler sur le continuum. Il est néanmoins intéressant de les étudier pour comprendre quels sont les GP qui se prêtent à ce rôle contrasté.

7.2.2.2 Profils plus homogènes

On trouve un deuxième type de GP, qui cette fois dépend de verbes dont le degré de productivité relative est lui même médian. Leur profil est plus homogène.

Le couple ($avec, curiosité_D$) relève de ce cas de figure dans le corpus FRANTEXT20 (cf. figure 7.1) : plus de la moitié des verbes auxquels il se rattache a un taux de sélection

moyen pour la préposition *avec*, et on voit se dessiner une gamme de verbes sémantiquement proches (*contempler*, *examiner*, *regarder*, *observer*, *considérer*, *fixer*).

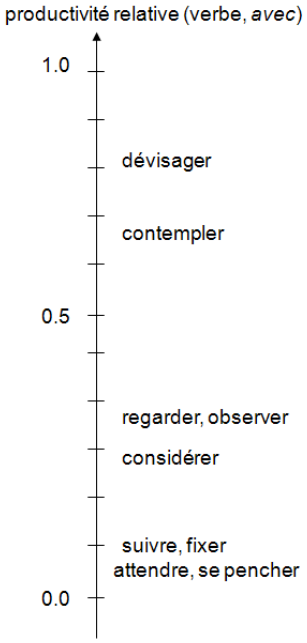

FIGURE 7.1 – Productivité relative des verbes associés au GP *avec curiosité*

Ces associations régulières ne sont pas de nature argumentale. L'application de tests linguistiques montre que les différents diagnostics concordent : ces compléments sont optionnels, ils acceptent le déplacement et l'extraction. Il s'agit clairement de compléments de manière, lesquels (si l'on fait abstraction des dépendants nominaux spécifiques) peuvent potentiellement s'adjoindre à la plupart des verbes. Mais les verbes qui figurent dans cette liste ont la caractéristique de déclencher préférentiellement la réalisation de ce type de complément. Ce type d'observation permet ainsi de déterminer un *pattern* sémantique qui semble bien établi dans le corpus, à savoir : VERBE D'OBSERVATION + COMPLÉMENT DE MANIÈRE. Plusieurs noms intègrent en effet ce patron aux côtés de *curiosité* : *attention*, *stupeur*, *étonnement*, *insistance*, *impertinence*.

Dans un cas comme celui-ci, l'observation des positions médianes permet donc de repérer des comportements véritablement intermédiaires entre arguments et ajouts, puisqu'il s'agit de *patterns*, d'associations sémantiques préférentielles entre certains types de GP et certains verbes. On quitte le domaine de la syntaxe, on est en-deçà de la relation argumentale puisque le complément n'est pas requis, il n'apparaît pas toujours, et les tests montreraient qu'il possède certaines propriétés des circonstants – le déplacement et l'extraction sont possibles. L'examen systématique de cette zone reste à faire. D'autres séries de GP semblent présenter ce statut intermédiaire, ainsi par exemple certains GP en *de* (*de bonheur*, *d'indignation*).

7.3 Profils de prépositions

On peut distinguer deux grands types de profils de prépositions :

- celles qui se cantonnent aux valeurs d'autonomie forte : *devant*, *derrière*, *après,* etc. Autrement dit, on ne les trouve associées qu'à des GP autonomes ;
- celles qui occupent au contraire un (plus ou moins) large spectre de valeurs d'autonomie. C'est particulièrement le cas des prépositions *à*, *de*, *dans*, *sur,* et, dans une moindre mesure, *avec* et *en.*

Au sein de cette deuxième catégorie, on peut regrouper les prépositions deux à deux, comme le montrent les figures 7.2 à 7.4, reproduites de (Fabre et Bourigault, 2008), qui montrent l'histogramme des valeurs d'autonomie des GP pour ces différentes prépositions sur le corpus LM10.

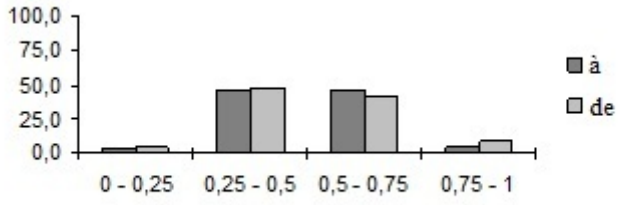

FIGURE 7.2 – Valeurs d'autonomie des GP introduits par les prépositions *à* et *de*

À et de Comme on l'a vu précédemment, les deux prépositions sont celles qui occupent la plus large gamme de valeurs : compléments très cohésifs (*à l'exigence*, *d'un voile*) jusqu'à des groupes très autonomes (*de front*, *à plat*), en passant par des situations intermédiaires et contrastées (*de bonheur*, *à la maison*).

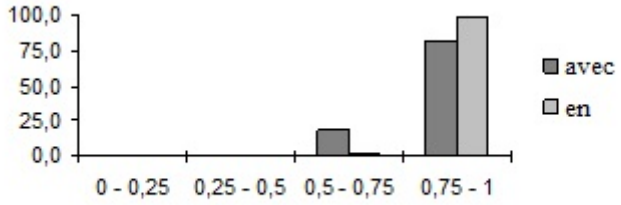

FIGURE 7.3 – Valeurs d'autonomie des GP introduits par les prépositions *avec* et *en*

Avec / En Quelques GP introduits par *en* et *avec* occupent des valeurs médianes d'autonomie. Dans le cas de *avec*, les valeurs moins hautes (autour de 0,6) correspondent prin-

cipalement, dans le corpus LM10, à l'emploi de *avec* marquant une relation entre entités (les dépendants sont en effet : *Israël*, *Etats-Unis*, *pays*, *gouvernement*, *représentant*, etc.). Il serait intéressant de regarder plus en détail comment les valeurs d'autonomie rendent compte du « déploiement de l'éventail » des valeurs de *avec* dont parle (Cadiot, 1997, p. 141 sq.) : si la valeur comitative que l'on vient de signaler se situe plutôt dans les valeurs les plus basses, la valeur de manière semble occuper plus particulièrement la zone haute (*avec précaution* ou *avec la rigueur* ont une valeur d'autonomie qui dépasse 0,9) (toujours dans le corpus LM10).

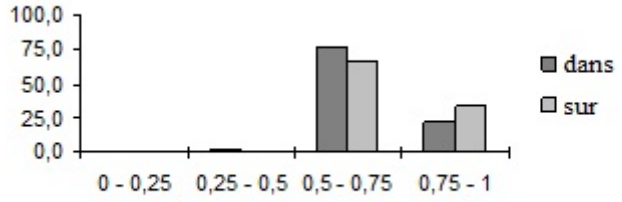

FIGURE 7.4 – Valeurs d'autonomie des GP introduits par les prépositions *dans* et *sur*

Dans / Sur Les GP en *dans* et en *sur* occupent principalement des valeurs d'autonomie haute et moyenne. La variation dans les valeurs d'autonomie est particulièrement facile à interpréter dans le cas du corpus FRANTEXT20, comme le montrent les exemples du tableau 7.6.

auton < 0,6	0,6< *auton* <0,8	*auton* > 0,8
labyrinthe	*château*	*jeunesse*
oreiller	*métro*	*espoir*
oubli	*nuit*	*style*

TABLE 7.6 – Diversité des GP en *dans*

Les GP dont l'autonomie se situe autour de 0,6 marquent principalement l'intériorité dans un espace clos –concret ou figuré. Plusieurs verbes associés ont une productivité relative haute avec cette préposition : *plonger*, *se jeter*, *pénétrer*, *entrer*... (Vaguer, 2005). Ceux qui se situent sur le pôle d'autonomie forte ne marquent généralement pas de localisation spatiale (mais expriment la temporalité, la manière). On trouve entre les deux des noms qui dénotent des lieux, sans que le rapport d'intériorité soit fréquemment marqué. Ces GP (*dans le château*, *dans le métro*, etc.) se combinent avec des verbes dont la valeur de $prod_R$ est non négligeable avec *dans* (*travailler*, *naître*, *rencontrer*, *installer*) – autrement dit, ces verbes s'accompagnent assez régulièrement d'une information spatiale. Le mot *nuit* illustre le cas d'un nom dont la sémantique est compatible avec ce double fonctionnement. Ainsi, *dans la nuit* peut donner lieu à la fois à une interprétation spatiale (la nuit est

alors considérée comme un espace, un milieu), et temporelle, comme l'ont montré A. Le Draoulec et D. Vigier [4].

7.4 Approche contrastive

Nous avons jusqu'ici commenté indifféremment les données des deux corpus LM10 et FRANTEXT20 pour illustrer les fonctionnements mis au jour par les mesures $prod_R$ et *auton*. Dans un second temps se pose la question de la généricité ou de la spécificité des profils dégagés selon les deux corpus. Il s'agit là de comparer les fonctionnements de la complémentation verbale dans deux types de discours, discours littéraire romanesque, et discours journalistique considéré dans toute sa diversité (puisque toutes les rubriques du quotidien ont été versées dans le corpus). J'indique ici les premiers éléments d'une étude contrastive en cours.

Profils des verbes Lorsque le verbe se construit avec une préposition unique ou qui prédomine, c'est presque systématiquement la même dans les deux corpus. Les différences sont peu nombreuses, et elles portent sur des associations peu productives, par exemple : *incorporer dans* (FRANTEXT20) / *incorporer à* (LM10). Lorsqu'il y a concurrence entre prépositions, on note des cas de stabilité remarquable d'un corpus à l'autre, mais également des cas de variation importante. Ainsi, pour le verbe *travailler* (tableau 7.7), la construction *travailler sur* (*un chantier*, *un projet*, *un thème*, etc.), s'impose dans le corpus LM10 aux côtés de la construction *travailler à*, alors que ses occurrences sont rares dans le corpus FRANTEXT [5].

travailler		
	LM10	FRANTEXT
sur	**0,19**	**(0,05)**
à	**0,19**	**0,3**
dans	0,16	0,2
avec	0,16	0,1
pour	0,1	0,1
comme, chez, jusqu'à, pendant, sous ...	$<$ 0,1 chacune	$<$ 0,1 chacune

TABLE 7.7 – Prépositions associées à *travailler* dans les deux corpus, selon la valeur de $prod_R$

4. Dans une communication intitulée « *Dans la nuit je suis descendu au village* : espace et temps dans quelques constructions en *dans* suivi d'un nom de partie de la journée » dans le cadre de la journée organisée par l'axe S'caladis (CLLE-ERSS) et consacrée à la localisation dans l'espace, 17 juin 2010, Toulouse.

5. Les chiffres de productivité relative ne correspondent pas tout à fait aux chiffres de productivité que nous avons présentés pour ce même verbe section 3.3.1, p. 59. Cela est dû au fait que l'on introduit ici une contrainte sur la fréquence des triplets pris en considération dans le calcul. Ce décalage n'invalide pas la remarque concernant le contraste de fonctionnement du verbe dans les deux corpus.

On trouve d'autres cas de verbes dont l'une des constructions s'impose nettement dans LM10, alors qu'elle est concurrencée par d'autres dans FRANTEXT. Il en va ainsi du verbe *crier* (*crier à* : *au sacrilège*, *à l'hérésie*, *au scandale*, etc.), ou du verbe *s'inscrire* (*s'inscrire dans* : *une problématique*, *une stratégie*, *un cadre*).

Profils des GP et des prépositions La valeur moyenne d'autonomie des prépositions est stable d'un corpus à l'autre : les histogrammes qui se dessinent pour le corpus FRANTEXT20 sont très proches de ceux que nous avons montrés pour LM10. On peut néanmoins observer quelques différences, qui vont toutes dans le même sens, à savoir, pour LM10, le décalage de quelques prépositions vers les valeurs d'autonomie inférieures. C'est le cas en particulier de la préposition *comme* :

On trouve dans LM10 une série de verbes dont la productivité relative est importante avec la préposition *comme* (cf. tableau 7.8).

LM10			FRANTEXT20		
verbe	prod	prod_R	verbe	prod	prod_R
se considérer	13	1	*considérer*	116	0,64
considérer	529	0,91	*apparaître*	55	0,2
se définir	17	0,81	*devenir*	27	0,19
se affirmer	11	0,73	*agir*	16	0,18
percevoir	79	0,62	*briller*	28	0,16
interpréter	37	0,61	*traiter*	22	0,15
sonner	17	0,59	*aimer*	17	0,13
ressentir	15	0,48	*flotter*	16	0,12
se comporter	12	0,48	*se ouvrir*	13	0,11
désigner	47	0,46	*battre*	13	0,1
apparaître	173	0,39	*éclater*	11	0,09

TABLE 7.8 – Valeurs de productivité pour la préposition *comme* dans les deux corpus

Alors que dans FRANTEXT20, la préposition marque presque uniquement la comparaison, et par conséquent une information périphérique pour les verbes concernés (*agir*, *briller*, *flotter*, etc.), LM10 comprend une grande série de verbes qui entretiennent une relation plus étroite avec la préposition. Les GP sont également très différents. On voit émerger de LM10 deux configurations :

- des GP de structure *comme* + <humain> (*comme directeur*, *comme chef*, *comme candidat*, *comme président*) ;
- des GP présentant une structure locutionnelle *comme* + dét + <abstrait> (de) (*comme le lieu, comme une sorte, comme un moyen, comme un coup*).

Alors que dans FRANTEXT20 dominent les GP de structure *comme* + dét + <humain, animal, objet> (*comme un chien, une fleur, une femme, un enfant, la pierre*).

Il serait intéressant de systématiser la comparaison, d'un corpus à l'autre, des fonctionnements sémantiques des prépositions avec les outils dont nous disposons. Par exemple, nous avons montré plus haut pour la préposition *dans* que les cas de GP moins autonomes marquent une relation d'intériorité (*dans le labyrinthe*, *le gouffre*, *le flot*, *la profondeur*), liée à l'emploi des verbes *s'enfoncer*, *sombrer*, *plonger*, etc. Dans le corpus LM10, les GP correspondants – c'est-à-dire les moins autonomes (*dans la création, la crise, la culture, la production*) sont associés à un autre type de vocabulaire verbal (*impliquer*, *spécialiser*, *se lancer*, *investir*, etc.). En s'intéressant aux GP les plus autonomes, on peut également cartographier l'information circonstancielle dans les deux corpus.

7.5 Conclusion

En examinant les propriétés générales des compléments verbaux dans deux grands corpus, nous adoptons un point de vue inédit sur des objets dont certains – les groupes prépositionnels – sont par ailleurs rarement examinés en tant que tels. Nous avons montré les potentialités de cette approche pour la description du fonctionnement des verbes, des prépositions, des GP : la masse d'informations qui se trouve synthétisée permet de dégager des tendances générales et des comportements différenciés d'un corpus à l'autre. Elle permet de mettre au jour des associations sémantiques qui peuvent alimenter les travaux descriptifs sur les prépositions.

Une cartographie systématique resterait à faire, qui permettrait de comparer l'information circonstancielle d'un corpus à l'autre, ou de comparer le fonctionnement d'une préposition à la manière de ce que nous avons esquissé pour *comme* et *dans*. Il serait intéressant de confronter ces observations aux descriptions qui ont été faites concernant la polysémie des prépositions (J.-C. Anscombre pour *sur* et *sous* (Anscombre, 1992), D. Leeman pour *dans* (Leeman, 1999), etc.). Il faudrait pour cela travailler au regroupement sémantique des GP, pour limiter l'éparpillement des observations sur des cas isolés. Dans l'immédiat, je m'intéresse plus particulièrement à la mise au jour de fonctionnements syntaxiques et sémantiques intermédiaires, qui se manifestent dans l'entre-deux entre les deux pôles d'autonomie et de cohésion maximales. Ce niveau intermédiaire est bien décrit par D. Legallois et J. François :

> (...) un « comportement » général incident aussi bien au niveau syntaxique qu'au niveau lexical, comportement de cooptation, de préférence mesurée statistiquement à partir de corpus, sans, encore une fois, qu'il soit redevable à quelque principe structural. (Legallois et François, 2006)

Un des intérêts de cette approche est en effet de fournir une mesure graduelle pour observer ces situations, comme nous avons commencé à l'illustrer pour le patron VERBE D'OBSERVATION + COMPLÉMENT DE MANIÈRE. La mesure d'autonomie fournit ainsi un instrument d'observation du continuum souvent évoqué entre arguments et circonstants.

Troisième partie

Bilan : Faire évoluer les liens entre linguistique et TAL

Présentation de la troisième partie

Dans cette partie de bilan, je présente la façon dont je considère l'apport mutuel de la linguistique et du TAL à la lumière des travaux que j'ai réalisés et auxquels j'ai participé, et depuis la position que j'occupe, à savoir celle d'une linguiste familière des méthodes et des objectifs du TAL. J'ai choisi de profiter de l'occasion qu'offre la rédaction d'un rapport d'habilitation pour dresser un bilan des évolutions actuelles de la collaboration entre linguistique et TAL, parce que mes contributions concernent ces différents aspects, mais aussi parce qu'il y a en la matière des enjeux de formation cruciaux, puisque nos enseignements visent à faire acquérir aux étudiants une double compétence en linguistique et en informatique. J'ai l'habitude de présenter aux étudiants deux volets de la collaboration entre linguistique et informatique : soit la linguistique se met au service d'applications informatiques, et on a alors affaire au domaine du TAL ; soit l'informatique se met au service de la linguistique, et le linguiste s'en sert pour augmenter ses capacités d'observation et d'analyse ; de ce mariage est née la linguistique outillée, et la linguistique de corpus. Mais cette double direction de coopération ne va pas de soi, et certains auteurs (cf. 8.1) considèrent au contraire que ces deux champs de recherche évoluent en restant très largement dans l'ignorance l'un de l'autre. Je consacre donc cette dernière partie à présenter la position qui est la mienne, et qui est plus généralement défendue par l'axe TAL du laboratoire CLLE, à savoir celle d'une linguistique qui prend véritablement part à la réflexion et aux travaux menés en TAL, et qui intègre les méthodes et les outils du TAL dans ses pratiques de recherche.

Le chapitre 8 s'interroge sur la première direction de coopération : soit la linguistique au service du TAL, et indique à quelles conditions cette collaboration peut se renforcer. Le chapitre 9 aborde l'autre versant, et montre ce que le TAL peut apporter à la linguistique en termes de méthodes, d'outils, de terrains d'investigation. La référence aux travaux exposés dans les deux premières parties me permet d'illustrer ces développements et de montrer comment ils s'inscrivent dans cette réflexion générale.

Chapitre 8

Contributions de la linguistique au TAL

Le travail du linguiste est généralement considéré en TAL comme une contribution au développement de ressources susceptibles de fournir aux systèmes les connaissances dont ils ont besoin pour traiter le langage de manière adéquate. D'où le développement de travaux de linguistique descriptive et formelle visant l'élaboration de lexiques, grammaires, corpus annotés (par exemple ceux décrits dans (Constant *et al.*, 2008)) dont l'objectif, envisagé comme une perspective à plus ou moins long terme, est d'intégrer ces systèmes pour en améliorer les performances, qu'il s'agisse de fournir des unités lexicales linguistiquement motivées pour les tâches de segmentation et d'indexation, d'en identifier les propriétés morpho-syntaxiques pour l'étiquetage et l'analyse syntaxique, d'établir des réseaux sémantiques pour étayer les procédures de catégorisation et de classification, etc.

Pourtant, l'apport de la linguistique aux évolutions actuelles du TAL est largement mis en question, et plusieurs auteurs, comme on va le voir, considèrent que le dialogue entre ces deux champs de recherche est devenu inexistant. Je pars de ce constat pour essayer dans un deuxième temps d'expliquer comment on peut travailler à améliorer cette interface.

8.1 Le TAL a-t-il encore besoin de la linguistique ?

On a assisté, depuis les années 1990, à un double mouvement défavorable à la linguistique : autour de la recherche d'information se sont développées des techniques qui se passent quasiment totalement de connaissances et de méthodes linguistiques. Les raisons principales tiennent aux exigences du traitement en temps réel, au caractère inadapté des ressources existantes, au coût d'intégration de ressources linguistiques nouvelles, et plus récemment au fait que la recherche sur le web rend largement superflues certaines solutions linguistiques sophistiquées, du fait du volume et de la redondance des données (« So why bother about variation, you find what you need anyway and you stay ignorant of what you do not find. » Vossen (2008, 24)). Parallèlement, les applications censées être plus gourmandes en connaissances linguistiques – extraction d'information, traduction automatique

– font appel à des approches quantitatives. (Habert et Zweigenbaum, 2002) et (Zweigenbaum *et al.*, 2008) résument très bien le changement, autour des années 1990, qui a vu les chantiers de modélisation linguistique, basés sur la constitution de règles de grammaire et de lexiques, perdre du terrain au profit de méthodes dirigées par les données. Alors que l'expertise linguistique était auparavant sollicitée pour produire des généralisations, désormais les modèles sont obtenus par apprentissage automatique. Dans ce cas, l'information linguistique se limite à celle dont on a besoin pour annoter les corpus d'apprentissage, ou pour valider (interpréter, affiner, filtrer) les règles apprises. Même si cette annotation peut intégrer des informations assez sophistiquées – comme dans le cas de l'extraction d'information (par exemple, la représentation de schémas d'événements entre entités) – il n'appartient pas au linguiste d'élaborer lui-même les généralisations, de construire les règles :

> « L'objectif d'une explicitation et d'un transfert d'expertise (celle du linguiste par exemple) est de plus en plus souvent remplacé par celui d'acquisition de connaissances linguistiques à partir de données. Les règles *a priori* cèdent la place aux régularités. La représentation des connaissances à l'acquisition des connaissances. » (Habert et Zweigenbaum, 2002, p. 86)

Plus récemment encore, S. Wintner fait le même diagnostic et suggère que la généralisation du recours aux corpus (et aux techniques d'apprentissage qui vont avec) a contribué à défaire le lien entre la linguistique et le TAL :

> « Frustrated by the inability of applications based on explicit linguistic knowledge to scale up to real-world needs, and, perhaps more deeply, frustrated with the dominating theories in formal linguistics, we looked instead to corpora that reflect language use as our sources of (implicit) knowledge. » (Wintner, 2009, p. 641)

Le TAL n'aurait plus besoin du rôle de médiateur du linguiste, puisqu'il s'agit de puiser directement à la source, et de faire l'impasse sur les connaissances linguistiques explicites pour exploiter les données brutes issues des corpus, seules susceptibles de donner accès à la réalité des usages langagiers que le TAL doit traiter. On en déduit tout simplement que les corpus remplacent les linguistes [1].

Ainsi, considérant, dans une perspective historique, le dialogue entre linguistique et informatique, K. Spärck-Jones estimait en 2007 qu'il était purement et simplement rompu, depuis longtemps :

> « Thus by the 1980s it was already clear that computational linguistics and natural language processing were advancing without referring significantly to mainstream linguistics or being significantly inadequate thereby. And this has become even more evident in the 1990s with the explosion of corpus-based research and the use of machine learning. » (Spärck-Jones, 2007).

1. On aurait pu s'attendre à ce que les corpus offrent au contraire un terrain de recherche commun, et enrichissent les échanges entre la linguistique et le TAL. Pourtant, ce sont ces mêmes corpus qui, en renouvelant les méthodes d'investigation linguistique, permettent aux linguistes de travailler sur des données d'usage riches et diversifiées, semblent contribuer par ailleurs à limiter les besoins du TAL en expertise linguistique...

De fait, nous dit Spärck-Jones, le domaine du TAL [2] se passe très bien des apports de la linguistique : les chercheurs en TAL se contentent généralement de mobiliser un bagage limité en linguistique, et n'ont guère besoin (pour entraîner leurs modèles) que de manipuler des notions assez élémentaires (catégories grammaticales, dépendances syntaxiques, classes sémantiques). Et de la même manière, les avancées de la linguistique *mainstream* ne se nourrissent pas de celles qui sont réalisées sur le versant computationnel ; D. Bourigault, poursuivant la réflexion de J.-C. Milner (1989), regrette « l'indifférence réciproque entre linguistes et techniciens de la langue », symptomatique d'une difficulté à concilier « proposition théorique » et « procédure technique » (Bourigault, 2007, p. 136).

D'où, au final, cette attaque virulente menée par l'éminente linguiste computationnelle vis-à-vis de la linguistique :

> « This is not to say that computational linguists can't, and shouldn't, take advantage of linguistics, or at least avoid culpable ignorance where linguists have something to offer. But the boot is now on the other foot, as Martin Kay was already claiming in 1973. He began his paper in the Rustin volume by saying that « For the most part, linguists are unaware of the importance that computers must one day have for their subject. » They still are, but that's no skin off our noses: We have more interesting things to do. » (Spärck-Jones, 2007, p. 440)

La situation que Spärck-Jones décrit est stérile pour la linguistique comme pour le TAL. Pour la linguistique, parce qu'elle se prive d'un champ d'application stimulant, et de l'opportunité de trouver dans le secteur des technologies du langage un terrain qui devrait naturellement lui inspirer de nouveaux objets d'étude et lui permettre d'appliquer et d'expérimenter ses propositions. Pour le TAL, parce qu'il n'est pas raisonnable de s'emparer d'objets linguistiques toujours plus complexes (les questions, l'expression de l'opinion, les inférences textuelles, le dialogue, la structure des discours, les genres de textes, etc.) en considérant que les procédures techniques sont tout et permettent de faire l'impasse sur une analyse approfondie de ces questions [3].

J'évoque ici les éléments qui permettent de nuancer ce constat, et de définir les conditions d'une collaboration possible entre TAL et linguistique. J'essaie de montrer le plus concrètement possible en quoi mes activités de recherche – mais aussi d'enseignement – contribuent à créer les conditions d'une réelle interface.

8.2 Créer les conditions d'une coopération

Lorsque K. Spärck-Jones déplore l'imperméabilité aux approches computationnelles d'une linguistique *mainstream*, et S. Wintner la frustration née des théories de la linguis-

2. Si l'on accepte de fondre dans ce terme, le seul qui fonctionne couramment en français, les deux versants « computational linguistics » et « natural language processing » auxquels K. Spärck-Jones fait allusion.

3. K. Spärck-Jones évoque en particulier de façon très éloquente la « seductive « look-no-hands » aura » des approches statistiques du langage.

tique formelle, elles désignent principalement là l'héritage chomskyen, qui a débouché sur des propositions très éloignées des approches légères et robustes, à large couverture et adaptées aux spécificités de textes variés et tout-venants que réclame le TAL. Cela ne signifie pas pour autant que la linguistique dans son ensemble se détourne de ce terrain applicatif : des possibilités de collaboration véritable existent, à condition de s'en donner les moyens. Pour commencer, il s'agit d'encourager la rencontre entre la linguistique et les méthodes et outils informatiques. Habert (2004, p.9) identifie « trois modalités du dialogue linguistique/informatique » :

- « mise d'une compétence d'informaticien au service d'un projet de linguistes »,
- « intégration de savoirs et savoir faire informatiques par des linguistes »,
- « collaboration finalisée avec des informaticiens dans un projet interdisciplinaire ».

J'ai expérimenté dans mes travaux les deux derniers volets de la collaboration entre les deux champs disciplinaires.

Tout d'abord, j'ai intégré à des degrés divers dans mes pratiques quotidiennes la maîtrise d'un outillage informatique qui me permet d'acquérir une certaine autonomie dans des tâches récurrentes comme le développement de programmes *perl* pour l'exploitation de grands corpus annotés en aval de SYNTEX ou l'utilisation d'un système de gestion de base de données pour manipuler les données une fois extraites. C'est en particulier ce à quoi j'ai eu recours dans le cadre du travail sur la distinction argument / circonstant pour organiser et exploiter les résultats des calculs de productivité et d'autonomie. C'est à cette condition que le linguiste peut appréhender comme il le souhaite et de façon autonome les résultats d'analyses automatiques qui sont souvent pléthoriques, et qu'il veut pouvoir aborder selon différents angles, différentes vues. Les étudiants que j'encadre, qui ont bénéficié d'une formation conséquente en informatique, possèdent souvent les habiletés techniques qui leur permettent d'acquérir le niveau d'autonomie nécessaire dans leurs travaux. Leurs compétences se diversifient. Dans le cadre de leur thèse, ils sont capables de traiter de grands volumes de données, d'y ajouter différentes couches d'annotation et de les exploiter. Ils savent travailler sur les données du web et développer des interfaces pour l'annotation et l'évaluation.

Par ailleurs, et pour l'essentiel, j'ai à la fois collaboré avec des collègues capables de mettre en œuvre des traitements informatiques qui ne sont pas à ma portée (D. Bourigault, N. Hathout, C. Jacquemin, P. Muller, L. Tanguy), et travaillé dans le cadre de projets collaboratifs avec des informaticiens – c'est le cas en particulier du projet d'acquisition de lexiques sémantiques (section 2.3), du projet ARIEL d'intégration de ressources linguistiques dans un programme de recherche d'informations (section 3.5), ou du projet ANNODIS pour la constitution d'un corpus annoté (section 4.2.1). Tous ces projets ont rassemblé des chercheurs issus d'équipes de linguistique et d'informatique, et les résultats obtenus sont le fruit de l'intégration des méthodes de ces deux champs. Lorsque ces situations de collaboration effective sont en place, on n'a alors pas de mal à déterminer comment des linguistes peuvent contribuer à l'avancée d'un projet de TAL.

8.2.1 Tâches linguistiques en TAL

Revenons une dernière fois à K. Spärck-Jones, qui nuance finalement son propos en expliquant qu'il y a malgré tout pour le TAL un danger à se couper de la linguistique (« We have to do computational linguistics and its congeners - natural language processing, etc., - properly »). Mais alors, nous dit-elle, ce dont le TAL a besoin, ce ne sont pas tant des modèles et des résultats récents de la linguistique que des garants méthodologiques qu'apporte une réflexion approfondie sur le langage.

Apports méthodologiques propres à la linguistique Ces garants méthodologiques tiennent à trois types d'exigence :

- être rigoureux en matière d'annotation des corpus : puisque ces annotations constituent les faits sur lesquels le traitement va reposer, de leur qualité dépend sa validité ;
- être prudent dans le maniement des intuitions sur le langage. Et, en la matière, on peut considérer que le linguiste est (malgré tout les pièges que l'on connaît) le mieux placé pour user avec méthode du jugement d'acceptabilité et, de façon générale, pour se livrer à un jugement réflexif sur la langue ;
- être scrupuleux dans la constitution des corpus, en se gardant néanmoins, nous dit Spärck-Jones, d'adopter une attitude puriste en la matière.

C'est au total, en apparence, peu de choses. Mais en réalité, c'est plaider pour que le matériau linguistique que le TAL traite et analyse soit considéré à chaque instant avec l'œil du linguiste, de manière à contrôler les conditions de l'expérience qui est menée, en termes de fiabilité des faits sur lesquels on travaille et de rigueur de l'évaluation. Le TAL a besoin du jugement linguistique à différents moments du développement d'un projet, et particulièrement en amont, dans la constitution des données, et en aval, dans la phase d'évaluation, moments critiques pour assurer la validité d'une démarche scientifique.

Dans le prolongement de cette idée, il me paraît utile de recenser brièvement les tâches dont on peut considérer qu'elles sont du ressort (pas nécessairement exclusif cependant) des linguistes dans un projet de TAL. Ce sont des tâches que j'ai en particulier été amenée à assurer dans le cadre des deux expériences de recherche, retracées dans le chapitre 2, au cours desquelles j'ai pleinement endossé le rôle d'une linguiste – linguiste au fait, dans les deux cas, des contraintes que font peser les développements informatiques sur la spécification des données linguistiques : le travail sur le repérage de variantes dérivationnelles avec C. Jacquemin (2.2) (cf. 2.2.4 pour une première évocation du rôle que j'y ai joué), et celui sur l'acquisition de lexiques sémantiques (2.3).

Délimitation de l'objet d'étude Le linguiste est sollicité en amont du traitement pour caractériser les phénomènes que l'on cherche à localiser dans les textes. Dans les deux expériences que j'évoque, il s'agissait de formuler des critères pour déterminer quel type de proximité sémantique on veut identifier. Ces critères ont été posés en termes de contraintes linguistiques : deux séquences de mots étaient considérées comme équivalentes si la même relation entre prédicat et argument pouvait être identifiée entre les deux variantes ; un nom et un verbe étaient considérés comme proches si une relation de type *qualia* pouvait

être identifiée entre eux. La linguistique intervient dans ce cas pour fournir les éléments de description formelle qui vont guider toute la procédure de découverte. Le travail dans un contexte appliqué amène aussi et surtout à travailler à la redélimitation des objets linguistiques, voire à l'invention de nouveaux objets de description, qui ne sont pas recensés parmi les notions déjà décrites par la linguistique, et le linguiste peut alors être amené à faire un long travail préalable de définition et de description – je pense par exemple à l'expression du dysfonctionnement technique dans des dialogues de travail (Vergely, 2004) ou aux segments d'obsolescence dans des documents encyclopédiques (Laignelet, 2009) qui ont fait l'objet de deux thèses au sein de l'équipe CLLE-ERSS.

Constitution du corpus Lorsqu'on conçoit un traitement visant l'acquisition de ressources lexicales se pose la question de l'adéquation du corpus à la tâche. Celle-ci peut se poser de diverses manières : quelle taille du corpus est requise, faut-il garantir une diversité ou au contraire une homogénéité interne, quelles caractéristiques formelles des textes sont à privilégier pour ne pas biaiser l'analyse ou pour augmenter les chances d'extraire de nombreuses occurrences du phénomène à traiter, etc. Dans le cas des deux expériences que je mentionne, la question très pragmatique de la disponibilité du corpus a largement prévalu sur ces considérations plus théoriques et méthodologiques. Néanmoins, dans le projet d'acquisition d'un lexique sémantique, nous avons opté pour un corpus de mode d'emploi parce que son vocabulaire présentait des propriétés intéressantes pour la tâche : abondance de noms concrets et de verbes décrivant les actions à effectuer sur les référents de ces noms. Le choix du corpus est une étape à ne pas négliger, et seule la familiarité avec des travaux de caractérisation situationnelle et linguistique des textes peut aider à l'objectiver (cf. (Habert, 2000) et (Biber *et al.*, 1998)). Ce qui suppose, en retour, de faire avancer la réflexion linguistique sur ces questions difficiles.

Construction de règles C'est la tâche traditionnellement dévolue au linguiste, qui sollicite ses compétences en matière de description et de modélisation. Dans le travail que nous avons mené sur les variantes dérivationnelles, les règles d'équivalence qui ont été implantées dans FASTER ont amélioré significativement les performances du système. Néanmoins, il est certain que cette tâche tend à être supplantée par les programmes d'apprentissage de manière à s'affranchir de cette phase, à la fois parce qu'elle est coûteuse, et parce que l'approche hypothético-déductive qui la sous-tend n'est pas nécessairement opérationnelle et peut-être remplacée, ou au moins assistée efficacement par des procédures de découverte. L'utilisation de méthodes d'apprentissage ne signifie pas pour autant l'élimination de l'expertise linguistique : le travail que nous avons réalisé sur la mise au jour de règles d'association sémantique entre noms et verbes à l'aide d'une méthode de programmation logique inductive, a nécessité au contraire de nombreux réglages et requis un accompagnement par l'expertise linguistique à tous les niveaux de traitement, et en premier lieu dans la définition des paramètres de l'apprentissage (paramètres contextuels et niveau d'étiquetage) (cf. section 2.3.2). Le recours à des méthodes d'apprentissage symbolique, aux vertus explicatives, dont les effets « boîte noire » sont fortement atténués, constitue un bon compromis, car elles offrent des procédures de découverte dont les résultats peuvent

être interprétés et amendés en prenant en compte de nouvelles hypothèses linguistiques, guidées par l'observation et l'expérimentation.

Annotation L'annotation de données est une étape requise pour pouvoir alimenter l'apprentissage (en exemples positifs et négatifs), et pour offrir une référence qui servira d'étalon pour l'évaluation des résultats obtenus par le système. Cette phase, cruciale, doit recevoir l'attention qu'elle mérite. Le recours systématique à la mesure d'accord inter-annotateurs permet de cadrer cette procédure. Une réflexion approfondie est requise pour déterminer la faisabilité de la tâche, les conditions dans lesquelles on peut la réaliser de façon à limiter les écueils du jugement réflexif, la manière dont on doit l'accompagner (protocole, guide d'annotation). A. Zaenen (2006), citée par (Spärck-Jones, 2007), met ainsi en garde contre la multiplication de tâches d'annotation qui ne sont pas faites dans les conditions optimales qui étaient celles du PENN TREEBANK, et qu'elle résume ainsi : définir une tâche linguistique sur laquelle l'accord inter-annotateur est élevé ; la confier à des étudiants avancés de linguistique, qui ont une bonne compréhension de la tâche à effectuer et sont supervisés par des experts ; enfin, consacrer à cette phase d'annotation le temps nécessaire. Cette réflexion a fortement progressé entre les deux expériences que nous avons retracées dans le deuxième chapitre, et celle que nous avons réalisée dans le cadre d'ANNODIS (cf. section 4.2.1), qui a consisté à mettre en place une véritable campagne d'annotation, comprenant l'écriture d'un guide, la formation d'annotateurs, des procédures de calcul d'accord inter-annotateurs, et qui prévoit dans la dernière phase du projet l'élaboration d'une annotation de référence sur une partie du corpus (Ho-Dac *et al.*, 2010).

Constitution de ressources La constitution de ressources – bases lexicales, grammaires – est considérée comme le cœur de l'activité linguistique en TAL. La tendance actuelle est néanmoins d'automatiser leur construction, en recyclant et combinant des ressources existantes et en mettant en place des procédures d'acquisition à partir de corpus – c'est le cas par exemple du développement d'un WordNet français, WOLF (Sagot *et al.*, 2009). Les linguistes sont principalement sollicités dans les phases d'évaluation et de validation. Différentes tâches peuvent cependant requérir le développement de ressources spécifiques. Dans l'expérience sur les variantes dérivationnelles, j'ai intégré un codage syntactico-sémantique léger dans le lexique de FASTER. Dans l'expérience sur les lexiques sémantiques, j'ai pris part à la définition d'une arborescence sémantique pour permettre d'effectuer un étiquetage sémantique du corpus (Bouillon *et al.*, 2001). Dans les deux cas, il s'agissait de ressources très modestes (un lexique de 1023 verbes et déverbaux, une liste de 33 étiquettes sémantiques hiérarchisées), qui avaient le mérite d'être totalement appropriées à la tâche, adaptées au corpus de traitement ainsi qu'aux méthodes et aux objectifs de l'apprentissage. Le travail entamé sur les voisins distributionnels montre que, même si la ressource est obtenue par des procédures automatiques, l'expertise linguistique est intégrée non pas simplement dans une phase finale de validation, mais à tout moment de la réflexion sur l'adéquation entre ressource et tâches (ici, recherche d'information et segmentation automatique), ce qui suppose une autre façon de considérer les étapes d'évaluation et de validation, qui constituent la dernière série de tâches linguistiques que j'évoque ici.

Interprétation et évaluation Si l'on travaille sur des procédures d'apprentissage, l'évaluation des traitements automatiques peut s'appuyer sur les annotations de référence qui ont été élaborées préalablement, (de préférence, nous dit A. Zaenen, par des linguistes) ; elle peut par ailleurs être réalisée en aval, en soumettant les résultats à des juges. Les linguistes sont souvent sollicités dans l'une ou l'autre phase. Le risque est alors grand de les voir cantonnés dans le rôle d'un évaluateur qui intervient à des moments bien déterminés d'un projet pour apporter sa caution, et se trouve soumis à des tâches d'évaluation complexes, aux consignes pas toujours bien définies (par exemple, valider des relations lexicales, déterminer des degrés de paraphrase, identifier les sens des mots). Certains projets de recherche permettent au contraire à l'analyse linguistique de jouer un rôle plus important, engageant une réflexion sur les conditions d'évaluation et sur la définition de la tâche elle-même.

Le travail autour des voisins distributionnels permet d'illustrer ce point : il s'agit de concevoir des modalités d'évaluation qui permettent d'éviter l'appel à des jugements artificiels (par exemple, un étiquetage manuel de la relation entre couples de voisins), pour encourager le recours à des tâches linguistiques plus réalistes, et plus contextualisées – comme le repérage de relations de discours ou la segmentation de textes. Ainsi, A. Tutin (2007) s'est intéressée aux capacités de la méthode d'analyse distributionnelle de mettre au jour l'organisation sémantique d'un lexique qu'elle connait bien, le lexique transdisciplinaire des écrits scientifiques. C'est sur cette base qu'elle peut alors repérer les limites du traitement, comme le caractère peu discriminant de certaines relations syntaxiques. Baroni et Lenci (2009) font ainsi la liste des nombreuses tâches qui permettent d'évaluer les systèmes d'analyse distributionnelle et qui correspondent à des procédures sinon totalement naturelles, du moins motivées linguistiquement (ex : jugements de similarité issus de tests de langue, tâches de catégorisation, définition de liens *qualia*, etc.).

De la même manière, nous avons vu, à propos de l'expérience d'intégration des voisins distributionnels dans un système de recherche d'information (3.5), que l'évaluation globale d'un système doit être complétée par une analyse circonstanciée des performances qui puisse prendre en compte les différents types de requêtes, voire de textes. Pour que des linguistes puissent contribuer à cette réflexion, il faut qu'ils puissent accéder au matériau langagier, ce qui suppose la mise en place de dispositifs d'évaluation plus élaborés.

8.2.2 Mise en place de projets collaboratifs

Comme on peut le constater à partir de ce recensement, les tâches dévolues au linguiste ou, en tous cas, nécessitant une contribution de nature linguistique sont, au moins potentiellement, nombreuses et cruciales. Mais c'est à condition de concevoir des modalités d'intervention et de collaboration adaptées, de manière à ce que la contribution des linguistes puisse être considérée de manière réaliste. Des projets de TAL associant de façon étroite linguistes et informaticiens constituent une bonne formule pour mettre en place des collaborations effectives. En effet, on ne peut pas espérer que la linguistique réalise de façon autonome des descriptions qui viendraient alimenter dans un second temps les travaux en TAL, dans la mesure où c'est précisément en rapport avec la tâche à effectuer, et par la

prise en compte des contraintes liées à l'automatisation, que les objets à décrire doivent être appréhendés (on a ainsi montré dans le premier chapitre de quelle façon le TAL amène à reconsidérer la notion d'équivalence sémantique). J'en donne quelques exemples tirés de mon expérience.

La collaboration de chercheurs de l'IRIT et de CLLE-ERSS dans le projet VOILADIS autour de la thèse de C. Adam permet à la fois de réaliser de façon effective une tâche automatique (la segmentation thématique), et de la faire évoluer en tenant compte des apports de la réflexion linguistique sur le discours – ce qui nous amène à la fois à prendre en compte la diversité des types de textes (Adam *et al.*, 2010), et à travailler sur des segments linguistiquement valides (travail à venir sur les structures énumératives issues du projet ANNODIS) (cf section 4.2.3).

Dans le projet ANNODIS, la constitution d'un corpus annoté discursivement est le résultat d'une collaboration entre linguistes et informaticiens qui a permis de mettre en place une chaîne d'annotation inscrite dans une plate-forme offrant les fonctionnalités requises, et en particulier capable de simuler deux modes de navigation dans les textes incluant une vue synoptique adaptée à l'approche macro (Ho-Dac *et al.*, 2009) (cf. section 4.2.1).

Enfin, le projet ARIEL, coordonné par L. Tanguy et J. Mothe (cf. section 3.5), consistait à travailler dans le cadre d'un dispositif combinant une plate-forme de tests et des bancs d'essai et offrant la possibilité de mener une évaluation plus fine et plus systématique de l'apport des méthodes et des ressources linguistiques en RI. La difficulté à évaluer cet apport tient en effet à la diversité des paramètres linguistiques à prendre en compte (que Mothe et Tanguy (2005) ont contribué à spécifier) et à la variété des expériences menées en termes de ressources mobilisées, de jeu de requêtes, de bases textuelles traitées. Le principe était d'offrir un environnement permettant de mieux contrôler la prise en compte d'informations linguistiques, à différents niveaux (morphologique, syntaxique, sémantique), et d'en mesurer systématiquement l'impact. En d'autres termes, il s'agissait de fournir une plate-forme d'évaluation la plus transparente possible. J'ai préalablement évoqué (section 3.5) les difficultés que nous avons néanmoins rencontrées dans cette expérience. En dernière instance, ce sont les modalités d'évaluation TREC qui font foi, dont les procédures restent assez opaques et critiquables. Nous avons évoqué dans (Picton *et al.*, 2008) le problème lié aux différents niveaux de requêtes (titre, description, narration) : l'évaluation de la pertinence a été faite sur la requête longue (<narr>), elle ne s'avère pas toujours valable pour la requête courte sur laquelle nous avons travaillé ; les critères de validation des textes pour une requête donnée sont totalement opaques, et aboutissent à des catégorisations parfois surprenantes. Cette première expérience plaide donc pour un travail au plus près de situations d'évaluation plus naturelles. Le projet ANR CAAS[4] qui démarre en octobre 2010 sous la responsabilité de Josiane Mothe (IRIT) est consacré à l'impact du contexte (nature des textes, types de requêtes, types d'usagers) sur la tâche de recherche d'information. Il s'appuiera sur des données recueillies dans des situations de besoin documentaire effectif, en particulier sur des *logs* de requêtes fournis par des partenaires industriels. Un travail

4. CAAS (ANR 2010-2013, programme contenus et interaction): Analyse Contextuelle et Recherche d'information Adaptative. Partenaires : IRIT (équipe SIG), CLLE-ERSS, LIA. Responsable : Josiane Mothe. C'est L. Tanguy qui coordonnera les activités de CLLE-ERSS.

de caractérisation linguistique des requêtes et des textes peut alors être mené dans des conditions d'expérimentation plus satisfaisantes.

8.2.3 Un contexte plus favorable ?

J'ai insisté sur les apports méthodologiques de la linguistique dans le TAL. Reste la question des problématiques linguistiques qui sont suscitées dans le champ du TAL, et de l'impact réel d'analyses linguistiques approfondies.

On a vu naître ces dernières années des applications qui prônent un accès plus précis et plus riche à l'information. Les campagnes TREC successives témoignent du raffinement des objectifs d'analyse. Les systèmes de questions-réponses remettent à l'ordre du jour des problématiques linguistiques fondamentales comme la paraphrase, l'inférence, la variation, et font renaître l'intérêt pour une indexation riche (sémantique) des textes. L'accès à l'information dans des domaines spécifiques ne peut s'appuyer sur des techniques à large spectre et nécessite le développement de connaissances ciblées. Les questions qui se posent sont à nouveau (partiellement) du ressort de la linguistique : comment évaluer la proximité sémantique de deux segments de textes ? Les traitements à effectuer sont-ils fonction des types de textes – et dans ce cas comment les caractériser et les prédire ? Les applications se diversifient, et de nouvelles problématiques apparaissent, dans lesquelles la linguistique est partie prenante (c'est le cas par exemple les travaux autour de l'expression du jugement). A. Nazarenko (2005) montre ainsi la richesse et la diversité des approches sémantiques mises en œuvre par les systèmes d'accès au contenu. Elle met en évidence l'intérêt d'une sémantique appliquée qui définit de nouveaux objets d'étude (avec, par exemple, la focalisation sur les entités nommées) et des traitements globaux, basés sur une grande diversité d'indices et de niveaux d'analyse.

Dans le cas d'applications ciblées à un domaine particulier, le développement de ressources linguistiques spécifiques constitue toujours un enjeu considérable, et sollicite le développement de données terminologiques (thésaurus, ontologies). Dans le cas des systèmes tout-terrain, il s'agit de faire intervenir des méthodes linguistiques de façon ciblée. Deux modalités sont bien identifiées (cf. par exemple (Katz et Lin, 2003)) :

- Intégration ciblée : les techniques qui se passent d'informations linguistiques restent au fondement des systèmes d'accès à l'information, le recours à des approches de TAL sophistiquées n'est envisagé que si leur apport en termes de performances est avéré : « recognize situations where linguistic techniques would help and employ them only when necessary. » (*ibid.*)
- Intégration différée : le recours à des techniques linguistiques est différé au moment où le système a limité le champ de recherche à un petit sous-ensemble de textes sélectionné par des techniques de recherche d'informations. Ainsi, dans les systèmes de questions-réponses, les traitements linguistiques interviennent dans la phase d'extraction de la réponse, après la phase de traitement des documents (on parle alors de « two-step paradigm »).

Par ailleurs, une utilisation plus massive de connaissances linguistiques s'avère probante dans certaines applications, comme l'a montré la société toulousaine Synapse dans le dé-

veloppement du moteur de questions-réponses QRISTAL, classé 1^{er} lors de la campagne EQueR de 2005 (Laurent *et al.*, 2006) : le texte est indexé en fonction des types de questions potentielles, ce qui amène à identifier des informations sémantiques relativement complexes, comme des liens de causalité, de but, de conséquence, des comparaisons, des opinions, etc. De nouvelles applications amènent à explorer de nouveaux objets. C'est le cas par exemple de l'analyse des opinions ou des sentiments, qui sollicitent, au moins potentiellement, les résultats de la linguistique de l'énonciation.

Cette focalisation sur l'analyse sémantique du contenu, sur les conditions d'énonciation, sur la prise en compte des variations du fonctionnement langagier (réflexion sur les types de textes dans le Web), appelle des questionnements linguistiques plus élaborés. Selon (Poibeau et Condamines, 2008, p. 7) :

> « On peut donc faire le pari d'un « retour du linguistique » pour l'accès à l'information, mais celui-ci ne pourra pas se faire sans une profonde réévaluation, voire une adaptation des méthodes de travail des linguistes eux-mêmes. »

La situation semble donc plus favorable aujourd'hui à cette réévaluation positive de l'apport de la linguistique dans les systèmes de TAL, dès lors que des linguistes s'impliquent dans les projets consacrés à l'exploration des nouveaux objets langagiers dont se saisit le TAL.

8.3 Conclusion : travailler à l'interface

Même si ces deux champs auraient toutes les raisons d'évoluer conjointement, il n'est pas réaliste de miser sur une rencontre naturelle et spontanée entre les recherches menées en linguistique et en TAL. Cette rencontre doit être suscitée et facilitée. Ce qui vient d'être dit plaide d'abord pour la formation de linguistes informaticiens, capables de bien maîtriser le fonctionnement des systèmes et de concevoir de façon autonome des procédures d'analyse : pour qu'une coopération entre linguistique et TAL soit possible, il est nécessaire que les linguistes qui s'y consacrent soient familiers des applications visées, maîtrisent certaines méthodes informatiques et soient mieux associés aux différentes étapes de la recherche, qu'il s'agisse de définir la tâche et l'objet d'étude, de déterminer les données à traiter, d'évaluer l'apport des outils. Il faut pour cela que la linguistique trouve un intérêt à se confronter aux outils, méthodes et ressources produites dans le champ du TAL. Le travail collaboratif au sein de projets doit faciliter cette démarche.

En retour, la linguistique a tout à gagner à s'imprégner de certaines méthodes développées en TAL. Et cette imprégnation facilitera les coopérations. Nous sommes plusieurs, au sein de CLLE-ERSS, à chercher à développer une linguistique alimentée par les méthodes du TAL. C'est le cas par exemple de Nabil Hathout en morphologie (Hathout, 2009), de Marie-Paule Péry-Woodley en discours (Péry-Woodley et Scott, 2006). C'est également ce que je cherche à faire concernant les objets de recherche qui m'occupent, à l'interface entre lexique et syntaxe, et entre lexique et discours. Le dernier chapitre fait le point sur les principales caractéristiques et les atouts d'une linguistique outillée par le TAL.

Chapitre 9

Intégrer le TAL dans les pratiques du linguiste

Ce chapitre s'intéresse à l'apport du TAL à la linguistique. Je montre comment la linguistique peut gagner à s'inspirer des apports méthodologiques du TAL, et à concevoir des dispositifs expérimentaux intégrant outils et procédures automatiques. Cette vision du TAL comme dispositif expérimental est au cœur des travaux que nous menons dans l'axe TAL, où nous cherchons à promouvoir de nouvelles procédures d'observation et d'analyse linguistiques. Mon objectif est ici de tenter d'alimenter cette réflexion commune, en montrant en particulier comment elle s'élabore dans les travaux que j'ai entrepris, souvent en collaboration avec ces collègues de l'axe TAL. Je fais le point sur la nature des outillages que j'ai manipulés dans mes travaux, et je montre comment ceux-ci m'amènent à positionner mes travaux dans le champ de la linguistique de corpus, en rapport avec l'opposition fondamentale entre analyses *corpus-based* et *corpus-driven*.

9.1 Linguistique outillée et TAL

La notion de linguistique outillée a été proposée par B. Habert dans le cadre d'un numéro de la revue RFLA rassemblant des « travaux de linguistes, de stylisticiens, de philologues qui mobilisent des outils de TAL pour leurs objectifs de recherche. » (Habert, 2004, p. 8). Cet outillage comprend les corpus enrichis d'annotation structurelles et linguistiques, les outils de programmation qui permettent de les explorer et de les analyser (Habert *et al.*, 1997), typiquement *perl* (Tanguy et Hathout, 2007), *python*, les environnements de programmation XML, mais également des outils bureautiques génériques comme les tableurs et les SGBD, qui permettent d'exploiter les résultats de l'extraction ou de l'annotation de données. Ces outils s'accompagnent également de méthodes. B. Habert mentionne la notion de patron et d'expression régulière, de similarité et de distance, auxquelles on peut ajouter celles de précision/rappel, et, de façon générale, les méthodes statistiques adaptées au traitement des données langagières qui peuvent être mis en œuvre à travers certains de ces outils.

9.1.1 Outils de base

Comme le dit B. Habert, la question de l'outillage pose celle de la formation des linguistes. Il s'agit de déterminer la part des compétences qui doivent s'enseigner au sein des seules filières TAL, et celles au contraire qui devraient irriguer l'ensemble d'une formation généraliste en sciences du langage et permettre de promouvoir une linguistique outillée pour tous, comme bagage commun d'un étudiant de linguistique – au même titre que les traitements statistiques entrent dans le bagage commun des étudiants de sociologie ou de psychologie. Dans ma pratique d'enseignante j'ai cherché avec mes collègues M.-P. Péry-Woodley et L. Tanguy à promouvoir, en marge de la filière TAL, des enseignements présentant cette approche outillée, et destinés à l'ensemble des étudiants de sciences du langage.

En L2, j'ai proposé un enseignement intégrant l'utilisation avancée de *Frantext* et du *Trésor de la Langue Française* dans sa version informatisée de manière à permettre aux étudiants d'amorcer le travail de constitution, de classement et de formalisation des observables à partir de faits linguistiques issus de textes ou d'énoncés lexicographiques ; en L3, j'ai mis en place avec M.-P. Péry-Woodley un cours sur les méthodes et les outils de la linguistique de corpus qui apprend aux étudiants à repérer des *patterns* réguliers, à dégager les faits qui permettent de faire la description de données qu'ils ont eux-mêmes collectées, à identifier les éléments de contraste majeurs entre deux corpus. La capacité à élaborer une description linguistique s'apprend dans cette confrontation aux corpus à travers les outils adéquats d'observation ; en master, j'ai mis en place avec L. Tanguy deux cours qui s'adressent à l'ensemble des étudiants et présente la diversité des données linguistiques dont dispose le linguiste, les principes de caractérisation et de documentation qui permettent de garantir leur validité et certains outils permettant de les manipuler – je propose quelques séances de cours sur les bases de données, dans l'esprit de (Habert, 2009). La capacité à mobiliser plusieurs types de données et de méthodes en faisant appel à des modes de collecte complémentaires (corpus, enquêtes, expériences) doit faire partie des compétences des linguistes.

Cette formation au travail sur les données linguistiques et aux méthodes de la linguistique de corpus me semble devoir constituer le bagage minimal d'un étudiant en linguistique. Un apprentissage conséquent de l'algorithmique et de la programmation devrait le compléter pour lui permettre de gagner en autonomie et de s'affranchir des spécificités des outils disponibles. Mais la formation en sciences du langage est ainsi faite à l'heure actuelle que ces enseignements relèvent presque exclusivement des filières consacrées au traitement automatique des langues, et il est très difficile d'amener des étudiants qui veulent faire, tout simplement, de la linguistique, et qui bénéficient de si peu d'heures d'enseignement dans les maquettes de leur formation, à opter pour cet apprentissage.

L'intégration de ces apprentissages techniques et méthodologiques dans les formations de linguistique est un enjeu important pour la discipline. Il s'agit d'être en mesure de faire une utilisation raisonnée des outils informatiques, de manière à choisir en connaissance de cause les ressources et les outils qu'il faut convoquer pour une recherche particulière. Par

exemple, pour une recherche lexicale, faut-il utiliser la base *Lexique*[1], le *TLFi*, faut-il faire des programmes d'extraction à partir de corpus, éventuellement à partir du Web ? Des compétences techniques sont requises pour parvenir à cette autonomie dans le traitement des données (Fabre et Lecolle, 2009; Habert, 2004). Une formation en statistiques est également indispensable.

9.1.2 Vers un outillage plus élaboré

Constituer son corpus selon ses propres critères, le nettoyer et le documenter, lui associer les couches d'annotation requises (étiquetage morpho-syntaxique, annotation structurelle, marquage d'informations linguistiques diverses), réaliser des extractions et lancer des procédures d'analyse : quantification des formes, études de répartition, recherche de corrélation entre plusieurs variables, etc. Le travail sur corpus requiert des compétences qui dépassent vite le bagage de l'utilisateur averti d'outils et de ressources déjà constituées (Tanguy et Hathout, 2007). L'idéal est alors d'être bien entouré, ce qui est mon cas, comme je l'ai signalé au début de la section 8.2.

J'évoque ici en quoi consiste véritablement selon moi un outillage avancé de l'analyse linguistique faisant appel à des méthodes de TAL. Je fais le point sur différentes procédures expérimentées dans le cadre de mes travaux.

Exploitation de corpus annotés syntaxiquement Comme je l'ai signalé précédemment (5.1), après avoir pris part au développement de l'analyseur SYNTEX, je l'ai ensuite utilisé comme support de l'analyse linguistique, de manière à travailler à partir non seulement des formes mais aussi des étiquettes catégorielles et des liens de dépendance (liens de rattachement prépositionnel) calculés par l'analyseur.

Il est clair que l'utilisation de SYNTEX comme instrument par des linguistes ne va pas de soi. Tout d'abord, les résultats comportent des erreurs d'analyse. S'il s'agit d'extraire et d'analyser des exemples particuliers relevant d'une construction particulière, le filtrage manuel permet d'y remédier. Mais lorsqu'on met au jour des tendances portant sur des masses importantes de données dont la validité ne peut pas être vérifiée, il n'est pas facile de savoir dans quelle mesure on peut s'accommoder d'une certaine marge d'erreur. Le maniement des sorties de SYNTEX nécessite une connaissance précise des mécanismes qui sont mis en œuvre dans l'analyseur, de manière à pouvoir comprendre et anticiper ces erreurs. Par ailleurs, les interfaces disponibles ne sont pas forcément adaptées. En particulier, l'utilisation de TIGERSEARCH requiert un apprentissage relativement lourd pour des linguistes qui n'ont pas l'habitude de faire appel à des langages de requête évolués. L'utilisation de petits programmes d'extraction est la solution la plus pratique. Enfin, et surtout, le fait de travailler en aval d'un analyseur amène à dépendre des décisions de traitement qui ont été réalisées, et qui risquent de produire des biais dans l'analyse. Ainsi, dans mon travail sur l'autonomie des GP, j'hérite de choix et de calculs qui concernent toutes les étapes de l'annotation. Si l'on prend l'exemple de la phase de segmentation des unités lexicales, une séquence prép + dét + N (*au contraire*, *à l'avenir*, *à l'horizon*) pourra avoir été identifiée

1. `http://www.lexique.org/`

comme unité locutionnelle (étiquette ADVGP, cf. section 7.2) si elle a été listée dans le lexique sur lequel se base l'étiqueteur. Une fois reconnue comme locution, elle est exclue des traitements ultérieurs que j'effectue sur les données, et n'est pas prise en compte dans le calcul de la productivité (cf. section 6.3). J'ai fait le choix de ne pas remettre en cause l'étiquetage des groupes prépositionnels, mais celui-ci pourrait être revu pour intégrer ces unités dans l'étude et vérifier ainsi le statut des GP qui ont été étiquetés.

Ces aspects de méthode sont importants car ils conditionnent l'utilisabilité en linguistique des résultats produits par les outils de TAL. Si l'on est conscient de ces aspects du traitement (par exemple, dans mon cas, des critères qui déterminent le rattachement des GP à un gouverneur ou leur statut d'entité flottante), on peut alors caractériser au mieux les données sur lesquelles on travaille, et profiter du gain offert par l'outil en termes de généricité des analyses, comme on a pu le voir avec les éléments d'observation que l'on recueille sur le profil général des verbes, des prépositions et des GP sur l'intégralité d'un corpus. En considérant toujours qu'il s'agit de résultats partiellement bruités, comme tout matériau que produit l'analyse automatique.

Marquage automatique des corpus Une autre méthode d'enrichissement des corpus consiste non plus à s'appuyer sur les annotations posées par un étiqueteur ou un parseur, mais à projeter dans les textes des éléments d'informations spécifiques, que l'on va ensuite incorporer dans l'analyse. Cette démarche a été mise en œuvre à deux occasions :

- à partir des voisins distributionnels : les thèses de C. Adam et F. Morlane-Hondère s'appuient sur la projection dans les textes des voisins produits par UPERY (cf. chapitre 4). Cette projection peut se faire selon différentes modalités (prise en compte ou pas de la relation syntaxique pour les voisins prédicats, réglage des seuils et des types de contextes à considérer, etc.) et permet soit d'utiliser le texte comme filtre pour identifier les voisins qui y cooccurrent (et conforter ainsi le lien qu'ils entretiennent), soit d'utiliser les voisins pour repérer des fonctionnements discursifs particuliers (cohésion lexicale, relations de discours).
- dans le projet ANNODIS : une phase de prémarquage automatique a permis de projeter sur les textes des ensembles de traits susceptibles de servir d'indices pour l'annotation manuelle de structures énumératives (cf. section 4.2.1). Ces traits sont définis *a priori* (ex : GP en position initiale, connecteurs, ponctuations) et des grammaires locales permettent de les repérer automatiquement dans les textes. Cette procédure, mise au point par L.-M. Ho-Dac, nous a permis de fournir à l'annotateur des éléments de guidage, qui l'amènent à privilégier la présence de zones denses en marques pour repérer les éventuelles structures énumératives qui s'y trouvent.

Ces phases de prémarquage présentent également les risques inhérents aux procédures automatiques : les voisins sont identifiés par leur forme et l'on risque d'établir des connections entre deux emplois différents d'un même mot ; le prémarquage effectué dans ANNODIS est approximatif, certains traits étant difficiles à recenser de manière exhaustive ou à identifier à coup sûr. Ces erreurs sont partiellement compensées par le fait que les marques sont nombreuses, et se confortent les unes les autres.

Procédures assistant l'investigation linguistique Le travail sur l'autonomie des groupes prépositionnels m'a amenée à réfléchir aux moyens de transposer des tests linguistiques en procédures automatiques permettant de tester certaines propriétés des GP dans un corpus – degré de dépendance à la tête, optionalité, mobilité. Ce travail de transposition me semble particulièrement intéressant à mener, car il permet d'appuyer l'analyse sur des éléments de quantification. J'ai montré que cette transposition n'allait pas de soi. On sait que le jugement d'acceptabilité s'interroge sur le possible en langue, alors que les corpus ne permettent d'examiner que l'attesté. Certaines tentatives ne se sont d'ailleurs pas avérées entièrement probantes, par exemple celle portant sur les tests de mobilité des GP, qui, concernant des configurations assez contraignantes et pas toujours bien analysées par SYNTEX, ont livré trop peu de résultats pour permettre de conclure (cf. section (6.4.2).

Le travail sur l'antonymie (cf. section 4.1.2) fournit une autre illustration de l'utilisation de procédures de TAL pour expérimenter une hypothèse linguistique. Dans ce cas, l'hypothèse est celle d'un fonctionnement à la fois paradigmatique et syntagmatique de la relation d'antonymie. Pour la tester, nous avons cherché à extraire des couples de mots (candidats à la relation d'antonymie) en combinant un test de substituabilité – basé sur l'analyse distributionnelle automatique – et la recherche de constructions marquant l'opposition – basée sur le repérage de patrons antonymiques. Le travail sur corpus avec des méthodes de TAL fournit ainsi des moyens d'expérimenter une hypothèse qui est difficile à vérifier par d'autres moyens. Là encore, cette approche doit être vue comme une approximation de l'hypothèse. D'abord parce qu'on travaille sur un corpus particulier. Ensuite, parce que, comme nous l'avons signalé, l'analyse distributionnelle automatique présente des limites importantes pour tester la substituabilité : la fréquence des mots introduit un biais très important. De la même manière, il y a plusieurs façons d'examiner le fonctionnement de ces mots sur l'axe syntagmatique : doit-on considérer des patrons particuliers, comme nous l'avons fait, ou la simple cooccurrence ? Les résultats et nos conclusions sont susceptibles de changer nettement si l'on change les conditions d'expérience, ce qui doit précisément nous encourager à le faire.

Procédures d'évaluation Le TAL fait appel à une large gamme de procédures d'évaluation. Voici une synthèse de celles que j'ai été amenée à utiliser (j'indique entre parenthèses l'étude concernée) :

- évaluation par la tâche (à travers l'expansion de requêtes et la segmentation thématique pour mesurer l'impact des voisins distributionnels) ;
- évaluation par rapport à une *baseline* (à travers la comparaison avec une technique de cooccurrence simple, ou avec le test du chi2, pour apprécier l'apport de règles linguistiques pour le repérage de variantes ou l'acquisition de liens sémantiques entre noms et verbes – chapitre 2) ;
- évaluation par rapport à une ressource de référence, qu'il s'agisse de dictionnaires (DICOSYN dans l'étude portant sur l'antonymie ; le TLFi dans l'étude sur la distinction argument / circonstant) ou d'annotation manuelle (recours à l'expertise linguistique sur ce même projet – chapitre 4 et 6) ;

– évaluation de la faisabilité d'une tâche par calcul d'un accord inter-annotateurs (dans plusieurs projets : repérage de variantes, antonymie, ANNODIS).

Ces outils d'évaluation peuvent permettre de garantir l'apport et la validité d'un résultat. Néanmoins, la valeur de cette contribution doit être nuancée, sur au moins deux points :

– des modalités d'évaluation artificielles sont parfois mises en place en TAL, qui ont pour effet de fausser l'évaluation : pour disposer de bancs d'essai, on simule les conditions d'une tâche (de recherche documentaire par exemple), hors de son environnement naturel[2] (seul à même de susciter un besoin informationnel réaliste) ;
– dans le cas des études sur les voisins distributionnels comme dans celles sur la distinction argument / circonstant, le recours au *gold standard* n'a pas de sens, puisqu'il s'agit justement de faire émerger de nouvelles catégories, susceptibles de faire évoluer les ressources existantes (Poibeau et Messiant, 2008).

Il s'agit de concevoir de nouveaux modes d'évaluation dans ces deux cas de figure. La linguistique doit pouvoir prendre part à cette réflexion. Dans le cas du travail que nous menons autour de l'autonomie des GP, la validation de la méthode peut certes dépendre de son intégration dans un module d'analyse syntaxique automatique. Mais il me semble qu'elle est aussi liée à sa capacité à faire émerger des faits linguistiques intéressants, à fournir un observatoire pertinent des phénomènes de complémentation, ce qui est beaucoup plus difficile à mesurer.

9.2 Linguistique de corpus et TAL

Le recours à des procédures automatiques assistant l'analyse linguistique prend sa source dans le développement des travaux sur corpus, et sur la disponibilité de corpus variés et volumineux.

9.2.1 Evolutions en linguistique de corpus

En 1999 nous organisions, A. Condamines, M. Péry-Woodley et moi-même, un atelier intitulé « corpus et TAL : pour une réflexion méthodologique », dans le cadre de la conférence TALN (Condamines *et al.*, 1999). Le besoin se faisait sentir à l'époque d'encadrer sur le plan des objectifs et des méthodes le recours croissant aux corpus en TAL et en linguistique, d'encourager leur constitution raisonnée, de s'interroger sur les méthodes d'annotation et sur la façon dont les résultats obtenus à partir d'un corpus pouvaient prétendre à une certaine généralisation. Aujourd'hui, il ne serait plus envisageable d'organiser un atelier sur une thématique aussi générale, tant la réflexion a progressé sur ces questions. Trois volets permettaient d'organiser les articles que nous avions retenus : la question de la constitution des corpus, de leur annotation, et de leur exploitation. Je propose de refaire

2. L Van der Plas, à l'occasion d'une communication, exprimait sa déception vis-à-vis des bancs d'essai dont elle avait disposé pour travailler sur la tâche de questions-réponses : les juges chargés de sélectionner les textes adaptés à chaque requête avaient fortement tendance à vérifier si les mots de la requête s'y trouvaient. Ce sont en effet des conditions d'évaluation qui mettent en péril une recherche consacrée à l'exploitation de relations sémantiques lâches (associatives) pour l'expansion de requêtes, qui était le but de sa thèse (Van der Plas, 2008) ...

brièvement le point sur ces questions dix ans plus tard, à la lumière des travaux auxquels j'ai participés.

La constitution des corpus Qu'est-ce qu'un bon corpus ? Quels critères doit-on considérer pour s'assurer de son adéquation à l'objet d'étude ? Nous faisions le constat de la variabilité des critères selon l'objectif de recherche qui motive la démarche de collecte. De fait, ce constat perdure. La question de la disponibilité du corpus reste critique, puisque le français ne dispose pas de corpus de référence diversifié, et malgré des initiatives telles que l'ILF ou le CNRTL, les dépôts de corpus ne se sont pas généralisés.

Je dois avouer que la question de la taille des corpus a souvent prévalu, en ce qui me concerne, sur l'objectif d'homogénéité. L'approche quantifiée de la complémentation, et plus encore la démarche d'analyse distributionnelle, réclament des grands corpus. Le grand événement fut pour moi, sur cette période, le passage de corpus de centaines de milliers de mots à des corpus de centaines de millions de mots. Evidemment, la notion de grands corpus invite à la modestie. Traiter un corpus de 200 millions de mots en français était un objectif irréaliste lorsque j'ai démarré mes travaux en linguistique de corpus ; c'est devenu une réalité dont nous nous sommes réjouis, lorsque D. Bourigault a réussi à analyser avec SYNTEX, puis UPERY, 10 années du *Monde* (au milieu des années 2000). Aujourd'hui, ce volume semble devenu à nouveau insuffisant au regard des milliards de mots traités en anglais par des outils d'analyse distributionnelle (Turney, 2006). Mais surtout, à mesure que le travail d'analyse s'affine sur les données que nous extrayons de ces grands corpus, la question de la nature des textes analysés se pose de manière aiguë. Travailler en aval de l'analyse distributionnelle automatique sur des corpus plus petits mais bien caractérisés, comme l'a fait (Tutin, 2007), garantit l'interprétabilité des rapprochements effectués, mais affecte leur qualité. Travailler sur de grands corpus dont l'homogénéité en genre est douteuse (pour LM10) et l'hétérogénéité thématique certaine (pour LM10 et WIKIPEDIA) pose véritablement problème [3].

Je n'ai cependant pas privilégié le volume de données dans tous mes travaux : dans le cadre du projet ANNODIS, nous avons constitué le corpus selon des critères linguistiques relativement bien établis, puisqu'il s'est agi de sélectionner des textes présentant des caractéristiques intéressantes sur le plan de la structuration du discours : textes expositifs longs, avec une mise en forme matérielle exploitable, relevant de genres et de thématiques variés.

L'annotation des corpus Dans cette période, j'ai expérimenté une large gamme de modalités d'annotation, soit en exploitant des annotations existantes, soit en participant à la définition d'annotations spécifiques aux projets en cours :

- annotation automatique par le biais de l'étiquetage catégoriel et de l'analyse syntaxique ;
- annotation manuelle de structures énumératives et de chaînes topicales dans le cadre du projet ANNODIS, annotation de transcriptions de consultations médicales qui

3. Sans compter que, comme nous l'a fait remarquer J. Delofeu à l'issue d'une présentation de notre travail, travailler sur *Le Monde*, est-ce vraiment travailler sur le français (et que dire alors de *Wikipedia*) ?

constituent le corpus sur lequel nous travaillons dans le cadre du projet INTERMEDE[4]. Dans ce projet, nous avons défini des consignes de transcription adaptées aux analyses automatiques que nous voulions mener sur le corpus : identification des tours de parole, des chevauchements, des pauses, des ratures énonciatives, homogénéisation des marques typographiques et de la ponctuation (Vergely *et al.*, 2009).
- annotation assistée (ANNODIS), avec la mise en place, en amont de l'annotation manuelle, des techniques de prémarquage que j'ai présentées plus haut. La chaîne d'annotation développée dans le projet ANNODIS accorde une place importante à la plate-forme d'annotation GLOZZ, qui a été adaptée aux spécificités du modèle et des procédures d'annotation que nous avons définis dans le volet « ascendant » du projet.

L'exploitation des corpus En 10 ans, les méthodes de la linguistique de corpus ont gagné en diversité et en complexité, de deux manières :
- en intégrant des procédures automatiques, comme je l'ai indiqué dans la section précédente. L'examen de concordances sur corpus bruts est une phase toujours nécessaire, mais le recours à diverses couches d'annotation et à des modes d'extraction qui offrent des points de vue multiples sur les données ont renouvelé la manière d'appréhender les corpus ;
- en mettant au point des mesures statistiques nouvelles, adaptées à l'étude de différents objets linguistiques. Ainsi, les techniques de calcul d'associations lexicales ont été complétées par des mesures d'associations grammaticales, comme celles proposées par (Stefanowitsch et Gries, 2009) pour calculer non plus seulement des associations entre formes, mais des associations entre une forme et une construction (collostructions).

En 10 ans, le champ de la linguistique de corpus s'est donc considérablement enrichi de méthodes nouvelles d'observation et d'analyse. On mesure d'autant mieux la nécessité évoquée au début de ce chapitre de former les linguistes aux statistiques et à la programmation, comme le signale S. Gries dans un article à paraître :

> « Why is it that we corpus linguists look at something (language) that is completely based on distributional/frequency-based probabilistic data (...) but most of our curricula do not contain even a single course on statistical methods (...) And why is it that we corpus linguists often must retrieve complex patterns from gigabytes of messy data in various encodings and forms or organization but most of our curricula do not contain even a single course on basic programming skills or relational databases (...) »(Gries, 2011)

9.2.2 Linguistique *corpus-based* ou *corpus-driven*

La distinction entre une linguistique *corpus-based* et une linguistique *corpus-driven* a été posée par E. Tognini Bonelli (2001). Celle-ci oppose deux paradigmes. D'un côté, les

4. INTERMEDE - Relation médecin malade et inégalités de santé (IRESP 2009-2010). Partenaires : U558 INSERM (Toulouse), LISST UMR 5193 (Toulouse), laboratoire de santé publique et d'épidémiologie (Nantes), LERASS (Toulouse). Resp. Thierry Lang (INSERM).

travaux *corpus-based* utilisent les corpus pour obtenir des illustrations et des décomptes de catégories préétablies. De l'autre, les travaux *corpus-driven* placent les corpus au cœur du processus d'élaboration des catégories. Dans le premier cas, la modélisation des phénomènes langagiers est préalable à l'exploration des corpus, elle est seulement éventuellement amendée (marginalement) par la découverte de données déviantes ; dans le second, la modélisation est le résultat d'un examen systématique et sans *a priori* des corpus, qui bouscule des distinctions préalables et façonne de nouveaux découpages, de nouveaux concepts. Cette opposition recoupe donc en grande partie celle qui distingue les approches *top-down* et les approches *bottom-up* présentées par (Biber *et al.*, 2007, p. 13 et 14) : l'approche *top-down* s'appuie sur la définition préalable du modèle de description, l'approche *bottom-up* est fondée sur des méthodes de classification automatique permettant de faire émerger les unités de description. Les travaux qui relèvent le plus clairement du versant *corpus-driven* sont ceux qui ont été menés autour des unités phraséologiques en anglais, et qui tendent à montrer que la démarcation entre unités syntaxiques et lexicales n'est pas nette. Les patrons du langage « formulaïque » (*formulaic language*) en sont le dernier avatar. Je propose de replacer par rapport à cette opposition les principales démarches que j'ai adoptées dans mes travaux.

a) Approches plutôt *corpus-based* :

Apprentissage d'information lexicale à partir de corpus annotés (2.3) : Dans ce cas de figure, l'annotation manuelle est réalisée à partir de catégories préexistantes. J'ai suivi cette voie en annotant les couples nom-verbe qui exhibaient une relation télique ou agentive. La technique d'apprentissage symbolique qui a été utilisée dans ce projet permettait ensuite d'identifier les contextes d'apparition de ces couples, de manière à acquérir de nouveaux couples pour peupler le lexique.

Elaboration de règles de reconnaissance de variantes de termes : La découverte de variantes verbales de termes nominaux s'est appuyée sur la production de règles faisant appel à des connaissances linguistiques génériques relatives à la structure argumentale du nom et du verbe.

Repérage automatique de la distinction argument-circonstant : Partant de l'hypothèse que la distinction existe, et qu'on a affaire à deux catégories de GP possédant des propriétés discriminantes, ma démarche initiale a consisté à identifier une mesure permettant de rendre compte de cette distinction, en essayant de transposer certains tests linguistiques. C'est également l'optique des travaux de Merlo et Ferrer (2006), qui se sont appuyés sur un corpus annoté pour alimenter des procédures d'apprentissage.

Dans tous les cas, le classement de ces études en approches *corpus-based* devrait être nuancé, car les règles sont sans cesse ajustées et affinées en tenant compte des spécificités du corpus, ou des premiers résultats d'évaluation.

b) Approches plutôt *corpus-driven* :

Analyse distributionnelle automatique : On ne sait pas *a priori* exactement quelles informations sémantiques sont accessibles à travers les propriétés distributionnelles. On fait néanmoins l'hypothèse que celles-ci sont un bon tremplin pour obtenir des proximités sémantiques, qui, une fois mises au jour, doivent être caractérisées.

Calcul du degré d'autonomie des groupes prépositionnels : Renonçant à l'opposition binaire entre argument et circonstant, la mesure d'autonomie que nous avons mise au point permet d'explorer toute la gamme des rapports de cohésion entre le GP et le verbe, aboutissant à l'étude d'objets intermédiaires, situés sur un continuum.

En résumé, mes travaux ont globalement évolué depuis l'approche *corpus-based* représentée par les deux expériences décrites dans le chapitre 2, vers les études *corpus-driven* relatives à l'analyse distributionnelle et à l'autonomie des groupes prépositionnels. Néanmoins, cette évolution doit être nuancée.

J'ai maintenu le choix de recourir de façon systématique à plusieurs couches d'annotation automatique (segmentation, étiquetage, analyse en dépendance). Dans (Römer, 2005), l'annotation est présentée comme relevant du versant *corpus-based* : les corpus sont annotés manuellement sur la base de catégories préétablies, et servent alors uniquement à produire des décomptes. Même si l'annotation porte sur des niveaux de description plus

élémentaires, on a toujours affaire à des abstractions qui figent la description et empêchent des redécoupages plus pertinents que l'analyse du corpus aurait pu faire émerger. De fait, le point de vue défendu par U. Römer se justifie théoriquement ; mais pour qu'il soit tenable pratiquement, et que l'on puisse réellement se passer d'une couche d'annotations, il faut que le niveau des formes de surface offre suffisamment de prise pour étudier les phénomènes visés. C'est justement le cas des travaux sur la phraséologie, où l'on calcule des suites de formes selon des critères de fréquence et de répartition. L'accès au lemme n'est pas souhaitable puisqu'il écraserait des distinctions pertinentes. C'est le cas aussi des travaux sur la forme progressive menés dans l'ouvrage coordonné par U. Römer. Mais on risque de se priver de la possibilité d'examiner des phénomènes de plus haut niveau. On est de fait frappé de l'évolution du champ de la linguistique de corpus, et en particulier par l'intérêt pour des études sur le *formulaïc language* qui banissent toute couche d'annotation préalable et nous ramènent au niveau du n-gramme.

Si j'ai décidé pour ma part de travailler à partir d'une couche d'annotation automatique préalable, c'est tout d'abord parce que l'annotation fait accéder à des généralisations qui constituent un apport incontestable pour l'approche choisie. C'est ensuite parce que, connaissant bien les outils de traitement, je suis consciente des décisions qui ont été prises. Enfin, le volume de données est tel qu'on peut s'accommoder de certaines erreurs dans la phase de prétraitement, à condition de disposer de procédures pour les limiter en aval (en particulier, l'utilisation de seuils de fréquences). On peut également faire le choix de revenir sur certaines annotations (comme relâcher certaines décisions de segmentation relatives à des unités complexes). Il reste toutefois difficile d'apprécier l'impact sur des résultats quantitatifs de certaines erreurs ou de certaines décisions particulières prises par les outils.

Je suis très réceptive à la critique que font McEnery *et al.* (2006) de cette opposition tranchée entre linguistique *corpus-based* et *corpus-driven*, laquelle s'accompagne généralement d'une attitude implicite assez dédaigneuse vis-à-vis des tenants de l'approche *corpus-based*. T. McEnery et ses collègues suggèrent qu'il n'y aurait en réalité pas de différence de nature entre des chercheurs qui vont vérifier - et donc éventuellement réviser - des théories préexistantes à la lumière des corpus, et ceux qui s'en servent pour forger de nouvelles théories en mettant à l'épreuve les théories préexistantes. Il note en particulier que l'approche *corpus-driven* a tendance à récolter des masses impressionnantes de données, mais qu'au total seule une petite partie de l'ensemble est examinée en détail pour déboucher sur une modélisation. Ce processus inévitable d'écrémage limite la portée d'une démarche inductive systématique.

> « Given that it is difficult to totally reject and dismiss preconceived theory, and intuitions are indeed called upon in corpus-driven linguistics, it is safe to conclude that there is no real difference between the corpus-driven demand to re-examine pre-corpus theories in the new framewok and corpus-based linguists' practice of testing and revising such theories. » (*op. cit.* p.10)

Il me semble qu'on peut considérer que les tenants de la linguistique *corpus-driven* ont eu tendance à surévaluer la distinction entre les approches *corpus-based* et *corpus-driven*. Néanmoins, c'est bien dans une approche principalement *corpus-driven* que j'inscris la

plupart de mes travaux récents, sans pour autant me priver d'y injecter des connaissances complémentaires, par le recours à des outils d'annotation automatique ou à des procédures de prémarquage.

9.3 Conclusion

Les outils de TAL appliqués aux corpus offrent à la linguistique des instruments d'expérimentation qui permettent de mettre en œuvre une approche *bottom-up*, émergente, des faits langagiers. Ils permettent également d'enrichir l'analyse par différents niveaux d'annotation qui donnent les moyens de se dégager du niveau des formes de surface pour effectuer des regroupements et accéder à des généralisations. Les deux séries de travaux que j'ai présentés, relatifs à l'utilisation de l'analyse distributionnelle automatique pour la mise au jour de relations sémantiques, et à l'utilisation de mesures de productivité pour l'étude de la complémentation, relèvent bien de cette approche. Celle-ci débouche sur une vision non catégorique, gradiente, des faits langagiers, dans la mesure où elle les aborde sous l'angle de la quantification, et par le biais de l'observation des usages.

Conclusions et perspectives

Ce rapport m'a amenée à dégager les principales caractéristiques de mon activité de recherche :

1. Mes travaux se situent à l'interface entre la linguistique et le TAL, de deux manières : je suis intéressée par la dimension applicative du TAL, et je pratique une linguistique outillée par le TAL.
2. La dimension applicative du TAL m'intéresse à plusieurs titres : d'abord, en tant qu'enseignante, parce que nous formons des étudiants qui sont amenés à intégrer le champ des industries de la langue. Ensuite, parce que je considère que la linguistique doit accompagner au plus près le développement de systèmes qui prennent en compte des objets linguistiques complexes et intéressants (avec parfois le risque d'utiliser des méthodes qui ne le sont pas). Enfin, parce que je pense que la linguistique peut également renouveler ses objets d'étude sur ce terrain appliqué.
3. J'ai essayé de préciser en quoi consistait cet outillage : appliqué à des (grands) corpus enrichis par différents niveaux d'annotation, il consiste à mettre en place des procédures de quantification pour dégager les fonctionnements de différents objets linguistiques dans tout l'espace du corpus. Ces procédures sont simples (mesures de productivité, mesures de similarité), ce qui me permet d'en maîtriser les principes et, grâce à diverses collaborations, la mise en œuvre (traitements reproductibles sur différents corpus, sorties manipulables selon différents paramètres).
4. Cet outillage aboutit à la mise en place de dispositifs expérimentaux. Ainsi, pour ce qui concerne les deux principaux axes de recherche que j'ai présentés, les voisins distributionnels offrent un point d'entrée pour travailler sur la cohésion lexicale ou sur la diversité des relations de proximité sémantique. La mesure d'autonomie des groupes prépositionnels fournit des éléments pour travailler sur la complémentation verbale en corpus.
5. Le travail linguistique qui en découle relève d'une approche non catégorique, graduelle, des faits observés. L'observation des usages montre que les catégories existantes ne suffisent pas, ou qu'il y a un continuum entre des pôles. On détecte des tendances, des contrastes. Les typologies que l'on peut dégager comportent toujours des zones faites d'erreurs et d'inclassables. Les conditions d'expérimentation peuvent être réglées à l'infini (modulation des seuils de fréquence, des mesures de similarité, etc.). En retour, cette approche appelle donc la mise en place de procédures d'évalua-

tion et de validation nouvelles (elle ne peut par exemple pas comparer ses résultats à des ressources dites *gold standard*)

6. Mes activités d'enseignement et de recherche se rejoignent en particulier sur la conviction qu'une formation plus technique et méthodologique axée sur le traitement des corpus, et des données que l'on en extrait, permet d'acquérir l'autonomie nécessaire pour travailler pleinement dans cette direction – même si le travail en équipe avec des informaticiens reste la formule indispensable.

Dans le cadre général que je viens de rappeler se dessinent deux chantiers immédiats, en rapport avec des thèses et des projets en cours, et deux objectifs plus généraux.

9.4 Tirer parti des dispositifs expérimentaux en place

Mon objectif immédiat est d'accompagner le travail des doctorants dans la finalisation de leur projet au sein des dispositifs que j'ai présentés, sur les questions de la détection de relations sémantiques et de l'utilisation d'indices lexicaux pour identifier des structures discursives. La mise en place des conditions d'expérimentation prend du temps dans le processus d'élaboration d'une thèse. Ainsi, la démarche de F. Morlane-Hondère a consisté à combiner des procédures de repérage par patrons (avec les multiples réglages que cette tâche suppose) avec l'information relative au voisinage distributionnel, ainsi qu'à concevoir des conditions d'évaluation particulières (comprenant en particulier l'élaboration de questionnaires). Les expériences menées par C. Adam l'ont amenée à projeter les voisins sur les textes selon diverses modalités, et à tester l'apport de cette information dans la réalisation de plusieurs tâches, en particulier la segmentation thématique, qu'il a fallu implémenter avec le maximum de rigueur possible, en collaboration avec P. Muller. Les premiers résultats de ces travaux ont été présentés dans le chapitre 4. Deux pistes me semblent plus particulièrement intéressantes à suivre maintenant :

- évaluer l'apport du voisinage distributionnel dans la détection de différentes relations lexicales, et en particulier comparer le cas de l'antonymie, dont la dimension syntagmatique a commencé à être confirmée par l'expérience menée par F. Morlane-Hondère, et celui de l'hyperonymie et de la synonymie ;
- utiliser le corpus annoté dans le cadre du projet ANNODIS (qui permet de travailler sur des structures discursives valides, plutôt que sur les indices de segmentation sommaires utilisés dans la tâche de segmentation thématique) pour étudier les propriétés de cohésion lexicale des structures annotées, en utilisant toujours le voisinage distributionnel comme base du calcul de cohésion.

Je compte également poursuivre le travail entamé sur la complémentation verbale, en collaboration avec J. Rebeyrolle, de manière à mener une analyse linguistique descriptive plus approfondie, portant sur le fonctionnement des prépositions dans ces corpus, en introduisant l'approche comparée que j'ai commencé à esquisser. Ce travail descriptif permettra d'estimer l'intérêt des outils d'observation mis en place.

9.5 Développer la recherche sur le versant applicatif

Deux projets vont m'amener à investir plus directement le terrain applicatif, dans le domaine de la recherche d'information. La collaboration régulière avec l'équipe d'Orange labs de Lannion[5] a débouché sur le démarrage, en octobre 2009, de la thèse de Fanny Lalleman, dans le cadre d'un contrat CIFRE. Cette thèse porte sur l'amélioration des méthodes de désambiguïsation dans le contexte des applications dites de *search* d'Orange. Une partie du travail d'encadrement consiste à encourager une approche linguistique du problème : il s'agit de mieux spécifier cette tâche de désambiguïsation (en quoi peut-on parler d'ambiguïté) et de développer des méthodes basées sur une meilleure prise en compte de l'impact de la variation linguistique, telle qu'on peut l'observer dans les données disponibles – requêtes utilisateurs, corpus issus des flux de *news*. F. Lalleman est ainsi été amenée à réexaminer la question de l'ambiguïté en intégrant dans ses expérimentations des techniques de catégorisation des textes. Un des objectifs de ce début de thèse a été de chercher à comprendre ce qui conditionne la répartition des résultats d'une requête dans plusieurs catégories thématiques : est-ce ce véritablement lié à une ambiguïté des mots ? A une pluralité d'usages correspondant à des contextes d'emploi diversifiés ?

Cette orientation applicative va être confortée par ma participation au projet CAAS[6] coordonné par J. Mothe (cf. section 8.2.2, page 146), qui va nous amener à travailler sur un meilleur paramétrage de la tâche de recherche d'information en fonction des caractéristiques linguistiques des requêtes et des textes. La première étape en sera un travail d'analyse linguistique fondé sur un corpus de *vraies* requêtes.

9.6 Faire évoluer la formation en linguistique

J'ai voulu signaler dans ce mémoire le lien étroit qui existe entre formation et recherche. Il n'est pas unilatéral, et ne se limite pas au cas où la recherche nourrit, par ses projets et ses résultats, les contenus d'enseignement, même si cet aspect est central. Les liens sont multiples. J'en citerai deux.

Tout d'abord, les évolutions méthodologiques de la recherche en linguistique, liées au changement de nature des données langagières, nécessitent en retour des changements importants dans la façon de concevoir ce que doivent être les compétences d'un étudiant de linguistique. L'outillage informatique et statistique doit prendre une place plus importante dans sa formation. C'est à cette condition que la linguistique pourra être envisagée – dans la recherche, mais aussi dans le monde de l'entreprise – comme une discipline capable de traiter et d'analyser les données langagières là où elles se trouvent et dans l'état où elles se trouvent – c'est-à-dire souvent bruitées, hétérogènes, volumineuses.

Ensuite, l'enseignement, en master professionnel tout particulièrement, nous amène à accompagner les étudiants sur des terrains nouveaux, et à les investir, ce qui en retour peut

5. Depuis plusieurs années, elle accueille des étudiants du master ECIL et du master TAL pendant leur période de stage.

6. Projet ANR2010-2013, programme contenus et interaction : Analyse Contextuelle et Recherche d'information Adaptative. Partenaires : IRIT (équipe SIG), CLLE-ERSS, LIA. Responsable : Josiane Mothe.

contribuer à faire évoluer certaines problématiques de recherche. C'est le cas, en particulier, du travail que nous effectuons dans le cadre du master ECIL [7] (Ergonomie Cognitive et Ingénierie Linguistique), qui forme des professionnels dans le secteur des facteurs humains, spécialistes de l'intégration des technologies dans l'activité de travail. S'y ajoute une compétence particulière dans la prise en compte des aspects langagiers de l'activité. Chaque année, plusieurs étudiants issus de la formation TAL de sciences du langage intègrent cette formation et se dotent au final d'une triple compétence en linguistique, informatique et ergonomie. Notre rôle est de conforter cette offre de formation en travaillant à la définition des atouts dont peut faire état un linguiste doté de ces compétences. Cette orientation doit nous encourager par exemple à participer à des projets de TAL qui répondent à des besoins applicatifs mieux spécifiés, et font appel à des procédures d'évaluation plus réalistes.

Mon objectif est donc de faire vivre une formation linguistique qui intègre au mieux les compétences techniques et méthodologiques que j'ai rappelées, et, pour ce qui concerne l'orientation professionnelle du master, qui soit articulée à des besoins applicatifs qu'il faut contribuer à définir et à développer. Cela suppose de déjouer un certain nombre d'écueils, l'un d'eux tient au fait que nous nous adressons à des étudiants qui viennent en sciences du langage avec des motivations d'une toute autre nature, en partie parce que la linguistique n'est précisément pas identifiée comme une discipline qui peut amener vers ce type d'activités.

9.7 Etendre les champs d'intervention de la linguistique outillée

Le développement de dispositifs expérimentaux permet à la linguistique de questionner avec de nouveaux moyens les objets qui l'intéressent (structures discursives, types de compléments, relations sémantiques, etc.). Mais il est important aussi que la linguistique puisse se saisir de problématiques qu'elle n'élabore pas elle-même, et répondre à des questions qui lui sont adressées par d'autres. Le TAL est une source possible de questionnement, mais ce n'est pas la seule. La notion d'intervention linguistique, comme action sur le terrain, réponse à une demande sociale (en termes de diagnostic, expertise, recommandations) a été discutée et défendue en particulier par A. Condamines (2005) et I. Léglise (2000). Toutes deux soulignent que, si l'on excepte peut-être les champs de la sociolinguistique et de la didactique du langage, peu de linguistes considèrent que l'intervention figure parmi les missions de leur discipline (contrairement à l'intervention ergonomique, psychologique ou sociologique). A travers sa propre expérience de linguiste de terrain [8] (Léglise, 1997), I. Léglise insiste pourtant sur « [l]a spécificité et [l]a force » du linguiste par rapport à d'autres chercheurs travaillant sur des données langagières : elle définit très justement la

7. Master co-dirigé par M.-P. Péry-Woodley et Claudette Mariné (CLLE-LTC) et relevant à la fois des mentions psychologie et sciences du langage.

8. Le travail qu'elle décrit concerne l'analyse des traces langagières de l'activité d'un opérateur acoustique. Elle montre que le relevé de patrons lexico-syntaxiques récurrents permet de mettre en évidence la nature des tâches cognitives réalisées (diagnostic, hypothèse, prédiction), aboutissant à une réévaluation de l'activité de cet opérateur.

spécificité de son approche par l'attention qu'il porte à la « matérialité même » du langage, quand les autres disciplines ont tendance à aller plus « directement à l'interprétation des discours ».

Cette question de l'intervention est au cœur du master ECIL. Pour reprendre le terme d'I. Léglise, notre optique consiste à travailler sur « la matérialité » des échanges et documents langagiers impliqués dans les situations de travail, avec des outils facilitant leur analyse et leur exploitation. Cette réflexion, dans sa finalité, est proche de celle qu'a engagée au cours des vingt dernières années le groupe « langage et travail » (Pène *et al.*, 2001), où linguistes et sociologues étudient ensemble la place du langage dans les activités de travail. Mais elle se caractérise, sur le plan méthodologique, par l'intégration de méthodes outillées et de techniques de TAL pour assister l'analyse de ces écrits. Les besoins d'analyse qui se font jour dans le monde du travail ne s'accompagnent pas toujours de données importantes en volume, elles peuvent être mal adaptées à des traitements extensifs, et les questionnements qui sont formulés ne se traduisent pas toujours facilement en termes de procédures d'analyse usuelles, préidentifiées. D'où un travail conséquent de reformulation du besoin et d'adaptation des méthodes qui nous sont familières.

D'autres sources de questionnement sont également à explorer.

La première émane plus généralement du champ des sciences humaines, qui ont à traiter des corpus textuels. Dans le cadre du projet INTERMEDE (9.2.1), nous avons fait l'expérience de la collaboration avec des partenaires épidémiologistes, sociologues et psychologues autour d'un même matériau langagier, recueilli dans le cadre de consultations médicales réalisées par des médecins généralistes. La question posée émane donc de la sphère médico-sociale, elle est celle du rôle de la consultation médicale dans la construction des inégalités sociales face à la santé. Elle doit déboucher sur des conclusions qui peuvent se traduire en recommandations à l'usage des médecins. Nous[9] avons intégré ce partenariat en proposant une expertise linguistique portant spécifiquement sur l'analyse du corpus de transcriptions. Ce type de projet ouvre un champ d'analyse nouveau pour la linguistique outillée. Le projet nous amène en particulier à mettre au jour des corrélations entre des variables externes (sexe, âge, CSP) et certaines caractéristiques linguistiques. Celles-ci peuvent relever d'une analyse très surfacique des corpus (nombre de mots, fréquence de renouvellement des tours de parole) ou passer par le marquage de traits linguistiques dans les textes (marques de subjectivité, vocabulaire technique, indices de reformulation). L'intérêt d'une telle collaboration est de pouvoir préciser la spécificité de notre démarche par rapport à des approches alternatives du matériau langagier (approches plus qualitatives, approches guidées par des catégorisations thématiques, approches quantitatives liées au dépouillement de questionnaires associés à la consultation, approches lexicométriques).

Je termine en évoquant brièvement une dernière piste qui me semble indiquer une articulation intéressante entre les méthodes de linguistique outillée et la notion d'intervention linguistique ; c'est celle de la didactique des langues, et plus particulièrement du français langue maternelle. Je prends part à la rentrée 2010 aux enseignements recherche du nouveau master EFE – spécialité professeur des écoles, par le biais de quelques heures

9. Participants au projet dans CLLE-ERSS : Anne Condamines, Amélie Josselin-Leray, Josette Rebeyrolle, Ludovic Tanguy, Pascale Vergely.

d'enseignement, qui vont porter sur les données linguistiques et sur la nature phraséologique des unités lexicales (collocations, collostructions). Par ailleurs le projet se précise au sein de l'équipe CLLE-ERSS, à l'initiative de Claudine Garcia-Debanc, de travailler avec des méthodes de linguistique outillée pour analyser des corpus de productions écrites d'enfants, qui offrent désormais un matériau suffisamment conséquent pour rendre possibles des explorations combinant marquages et quantifications. Il s'agit alors de considérer ce matériau comme un moyen de mieux analyser les difficultés et les mécanismes d'apprentissage (concernant par exemple la maîtrise de la structuration du discours ou des structures syntaxiques), et d'élaborer, à destination des enseignants et en collaboration avec les collègues spécialistes du champ didactique (principe d'intervention), des propositions pour travailler de manière spécifique sur les difficultés mises au jour.

Index

Table des figures

Liste des tableaux

Bibliographie

Adam, C. et Morlane-Hondère, F. (2009). Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique. *In Rencontres des jeunes chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, Senlis, France.

Adam, C., Muller, P. et Fabre, C. (2010). Une évaluation de l'impact des types de textes sur la tâche de segmentation thématique. *In actes de la 13e conférence sur le traitement automatique des langues naturelles (TALN)*, Montréal, Canada.

Adam, C. et Vergez-Couret, M. (2010). Signalling Elaboration: Combining Gerund Clauses with Lexical Cues. *In Proceedings of MAD: Multidisciplinary Approaches to Discourse (MAD 2010)*, Moissac, France.

Anscombre, J.-C. (1992). Sur/Sous : de la localisation spatiale à la localisation temporelle. *Lexique*, 11:111–145.

Baroni, M. et Lenci, A. (2009). One distributional memory, many semantic spaces. *In Proceedings of the EACL Workshop on GEMS: Geometrical Models of Natural Language Semantics*, pages 1–8, Athènes, Grèce. Association for Computational Linguistics.

Barzilay, R. et McKeown, K. (2001). Extracting paraphrases from a parallel corpus. *In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, volume 39, pages 50–57, Toulouse, France.

Barzilay, R., McKeown, K. et Elhadad, M. (1999). Information Fusion in the Context of Multi-document Summarization. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, University of Maryland, Etats-Unis.

Basili, R., Pazienza, M. et Vindigni, M. (1997). *Corpus-driven unsupervised learning of verb subcategorization frames.* Springer-Verlag, Berlin, Allemagne.

Bestgen, Y. et Piérard, S. (2006). Comment évaluer les algorithmes de segmentation automatiques? essai de construction d'un matériel de référence. *In Verbum ex machina: actes de la 13e conférence sur le traitement automatique des langues naturelles (TALN)*.

Biber, D., Connor, U. et Upton, T. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins Publishing Co.

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

Bonami, O. (1999). *Les constructions du verbe: le cas des groupes prépositionnels argumentaux. Analyse syntaxique, sémantique et lexicale*. Thèse de doctorat, Université Paris 7.

Borillo, A. (1990). À propos de la localisation spatiale. *Langue française*, 86(1):75–84.

Bouaud, J., Habert, B., Nazarenko, A. et Zweigenbaum, P. (2000). Regroupements issus de dépendances syntaxiques sur un corpus de spécialité: catégorisation et confrontation à deux conceptualisations du domaine. *In Ingénierie des connaissances : évolutions récentes et nouveaux défis*, pages 275–290. Charlet, J., Zacklad, M., Kassel, G. et Bourigault, D. (eds).

Boughanem, M., Dkaki, T., Mothe, J. et Soulé-Dupuy, C. (1998). Mercure at TREC 7. *In* Harman, D., éditeur : *Proceedings of TREC 7: Text Retrieval Conference*, pages 355–360, Washington.

Bouillon, P., Claveau, V., Fabre, C. et Sébillot, P. (2001). Using Part-of-Speech and Semantic Tagging for the Corpus-Based Learning of Qualia Structure Elements. *In Proceedings of the First International Workshop on Generative Approaches to the Lexicon : GL'2001*, Genève, Suisse.

Bouillon, P., Fabre, C., Sébillot, P. et Jacqmin, L. (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL (traitement automatique des langues), numéro spécial Traitement automatique des langues pour la recherche d'information*, 41(2):367–393.

Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L. et Tiedemann, J. (2005). Linguistic Knowledge and Question Answering. *TAL (Traitement Automatique des Langues)*, 46(3):15–39.

Bourigault, D. (1994). *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*. Thèse de doctorat, Ecole des Hautes Etudes en Sciences Sociales, Paris.

Bourigault, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN)*, pages 75–84, Nancy.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel: Syntex.* Habilitation à diriger des recherches. Université Toulouse 2 - Le Mirail.

Bourigault, D. et Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 25:131–151.

Bourigault, D., Fabre, C., Frérot, C., Jacques, M. et Ozdowska, S. (2005). Syntex, analyseur syntaxique de corpus. *In Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, pages 17–20.

Bourigault, D. et Lame, G. (2002). Analyse distributionnelle et structuration de terminologie: Application à la construction d'une ontologie documentaire du Droit. *TAL (Traitement Automatique des Langues)*, 43(1):129–150.

Brent, M. (1993). From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

Cadiot, P. (1997). *Les prépositions abstraites en français.* Armand Colin.

Callison-Burch, C., Cohn, T. et Lapata, M. (2008). ParaMetric: An Automatic Evaluation Metric for Paraphrasing. *In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 97–104, Manchester.

Charolles, M. et Péry-Woodley, M.-P. (2005). Les adverbiaux cadratifs. *Langue Française*, 148:3–8.

Church, K. et Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Claveau, V. (2003). *Acquisition automatique de lexiques sémantiques pour la recherche d'information.* Thèse de doctorat, Université de Rennes I.

Claveau, V. et L'Homme, M. (2004). Discovering Specific Semantic Relationships between Nouns and Verbs in a Specialized French Corpus. *In Proceedings of the 3rd Workshop on Computational Terminology (CompuTerm'04)*, Genève.

Claveau, V., Sébillot, P. et de Beaulieu, C. (2004). Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. *In Actes de la 11ème conférence de Traitement automatique des langues naturelles (TALN)*, Fès, Maroc.

Condamines, A. (2000). Les bases théoriques du groupe toulousain" Sémantique et Corpus": ancrages et perspectives. *Cahiers de grammaire*, 25:5–28.

Condamines, A. (2005). Analyse linguistique de documents d'entreprises: demande appliquée et théorisation des phénomènes. *Production, compréhension et usages des écrits techniques au travail, Octares, Paris*, pages 17–30.

Condamines, A., Fabre, C. et Péry-Woodley, M.-P., éditeurs (1999). *Corpus et TAL : pour une réflexion méthodologique.* Atelier thématique, TALN'99, Cargese, France.

Constant, M., Dister, A., Emirkanian, L. et Piron, S., éditeurs (2008). *Description linguistique pour le traitement automatique du français*. Presses univ. de Louvain.

Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press.

Dagan, I., Glickman, O. et Magnini, B. (2006). The PASCAL recognising textual entailment challenge. *Lecture Notes in Computer Science*, 3944:177–190.

Dras, M. (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Thèse de doctorat, Macquarie University, Australia.

Dubois, J. et Dubois-Charlier, F. (1997). Synonymie syntaxique et classification des verbes français. *Langages*, 31:51–71.

Duclaye, F. (2003). *Apprentissage automatique de relations d'équivalence sémantique à partir du Web*. Thèse de doctorat, ENST - INFRES Informatique et Réseaux.

Duclaye, F., Yvon, F. et Collin, O. (2003). Learning paraphrases to improve a question-answering system. *In Proceedings of the EACL Workshop on Natural Language Processing for Question Answering*, pages 35–41, Budapest.

Fabre, C. (1996). *Interprétation automatique des séquences binominales en français et en anglais*. Thèse de doctorat, Université Rennes I.

Fabre, C. (1998). Repérage de variantes dérivationnelles de termes. *Carnets de grammaire. Rapports internes de CLLE-ERSS*, 3.

Fabre, C. et Bourigault, D. (2001). Linguistic clues for corpus-based acquisition of lexical dependencies. *In Proceedings of the Corpus Linguistic Conference*, pages 176–184, Lancaster.

Fabre, C. et Bourigault, D. (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. *In Verbum ex machina: actes de la 13e conférence sur le traitement automatique des langues naturelles (TALN)*, pages 121–130, Leuven.

Fabre, C. et Bourigault, D. (2008). Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. *Journal of French Language Studies*, 18(01):87–102.

Fabre, C. et Frérot, C. (2002). Groupes prépositionnels arguments ou circonstants: vers un repérage automatique en corpus. *In Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, pages 215–224, Nancy.

Fabre, C. et Habert, B. (1998). Acquisition de relations entre mots pour une lecture sémantique de corpus. *In Actes des 4e journées internationales d'analyse statistique des données textuelles (JADT)*, pages 273–282, Nice.

Fabre, C., Habert, B. et Labbé, D. (1997). La polysémie dans la langue générale et les discours spécialisés. *Sémiotiques*, 13:15–31.

Fabre, C. et Jacquemin, C. (2000). Boosting variant recognition with light semantics. *In Proceedings of the 18th conference on Computational linguistics (COLING)*, pages 264–270, Morristown, NJ, USA.

Fabre, C. et Kupść, A. (2009). Large and noisy vs. small and reliable: combining 2 types of corpora for adjective valence extraction. *In Proceedings of the 5th Corpus Linguistics Conference*, Liverpool.

Fabre, C. et Lecolle, M. (2009). S'approprier des instruments d'observation de la langue pour élaborer des recherches : le TLFi et Frantext pour des étudiants de linguistique. *Ecrits de savoir, M. Kara (coord.), Pratiques*, 143-144.:139–152.

Fabre, C., Rebeyrolle, J. et Ho-Dac, L. (2008). Examen du statut des syntagmes prépositionnels à la lumière de données issues de corpus annotés. *In Actes du 1er Congrès Mondial de Linguistique Française (CMLF)*, Paris.

Fabre, C. et Sébillot, P. (1999). Semantic Interpretation of Binominal Sequences and Information Retrieval. *In Proceedings of International ICSC Congress on Computational Intelligence: Methods and Applications,CIMA'99, Symposium on Advances in Intelligent Data Analysis AIDA'99*, Rochester, N.Y., USA.

Fellbaum, C. (1995). Co-occurrence and antonymy. *International Journal of Lexicography*, 8:281–303.

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G. et Jacquemin, C. (2001). Terminological variants for document selection and question/answer matching. *In Proceedings of the ACL-EACL Workshop on Open-Domain Question Answering*, pages 46–53, Toulouse.

Fort, K. et Guillaume, B. (2007). PrepLex: un lexique des prépositions du français pour l'analyse syntaxique. *In Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles*, pages 219–228, Toulouse.

Frérot, C. (2005). *Construction et évaluation en corpus variés de lexiques syntaxiques pour la résolution des ambiguités de rattachement prépositionnel.* Thèse de doctorat, Université Toulouse 2 - Le Mirail.

Frérot, C., Bourigault, D. et Fabre, C. (2003). Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus: Le cas du rattachement verbal à distance de la préposition "de". *TAL (Traitement automatique des langues)*, 44(3):167–186.

Fuchs, C. (1982). *La paraphrase.* Linguistique nouvelle. PUF, Paris.

Fuchs, C. (1994). *Paraphrase et énonciation.* L'homme dans la langue. Ophrys, Paris.

Galy, E. et Bourigault, D. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *In G. Williams, éd., Texte et corpus: Actes des 4e journées de linguistique de corpus (JLC)*, pages 163–174, Lorient.

Goldberg, A. et Ackerman, F. (2001). The pragmatics of obligatory adjuncts. *Language*, 77(4):798–814.

Grefenstette, G. (1994a). Corpus-Derived First, Second and Third-Order Word Affinities. *In Proceedings of the European Association for Lexicography Conference (EURALEX)*, Amsterdam.

Grefenstette, G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Dordrecht: Kluwer Academic Publishers.

Gries, S. (2011). Methodological skills in corpus linguistics: a polemic and some pointers towards quantitative methods. *à paraître. En ligne sur : http://www.linguistics.ucsb.edu/faculty/stgries/research/overview-research.html.*

Grimshaw, J. (1992). *Argument structure*. MIT Press, Cambridge, MA.

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

Habert, B. (2004). Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs. *Revue française de linguistique appliquée*, IX(1):5–24.

Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Ophrys, collection l'Essentiel français, Gap/Paris.

Habert, B. (2009). *Construire des bases de données pour le français*. Ophrys.

Habert, B. et Fabre, C. (1998). Relations de dépendance syntaxique et sémantique distributionnelle. *In* Mellet, S. et Vuillaume, M., éditeurs : *Mots chiffrés et déchiffrés. Mélanges offerts à Étienne Brunet*, Travaux de linguistique quantitative, pages 113–123. Honoré Champion, Paris.

Habert, B. et Fabre, C. (1999). Elementary dependency trees for identifying corpus-specific semantic classes. *Computers and the Humanities*, 33(3):207–219.

Habert, B., Fabre, C. et Issac, F. (1997). *De l'écrit au numérique (constituer, normaliser et exploiter les corpus électroniques)*. InterEditions.

Habert, B. et Nazarenko, A. (1996). La syntaxe comme marche-pied de l'acquisition des connaissances: bilan critique d'une expérience. *Journées sur l'acquisition des connaissances*, pages 137–142.

Habert, B. et Zweigenbaum, P. (2002). Régler les règles. *Traitement automatique des langues*, 43(3):83–105.

Habert, B. et Zweigenbaum, P. (2003). Classer les mots: sémantique à gros grain et méthodologie harrissienne. *Revue de Sémantique et Pragmatique*, 12:101–119.

Halliday, M. et Hasan, R. (1976). *Cohesion in English*. Longman London.

Harris, Z., Gottfried, M., Ryckman, T., Mattick Jr, P., Daladier, A., Harris, T. et Harris, S. (1989). *The form of information in science: analysis of an immunology sublanguage*. Kluwer Academic Publishers.

Hasan, R. (1984). *Understanding Reading Comprehension: Cognition, Language, and the Structure of Prose*, chapitre Coherence and cohesive harmony, pages 181–219. Newark: International Reading Association.

Hathout, N. (2009). *Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie*. Habilitation à Diriger des Recherches en Sciences du Langage, Université Toulouse 2 - Le Mirail.

Hathout, N. et Tanguy, L. (2005). Webaffix : une boîte à outils d'acquisition lexicale à partir du Web. *Revue Québéquoise de Linguistique*, 32:61–84.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *In Proceedings of 14th International Conference on Computational Linguistics (COLING)*, Nantes, France.

Hearst, M. A. (1997). TextTiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Hermann, E., Leblois, S., Mazeau, M., Bourigault, D., Fabre, C., Travadel, S., Durgeat, P. et Nouvel, D. (2008). Outils de Traitement Automatique des Langues appliqués aux comptes rendus d'incidents et d'accidents. *In Actes du 16e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement*, Avignon.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P. et Rebeyrolle, J. (2009). A top-down approach to discourse-level annotation. *In Proceedings of the Corpus Linguistics Conference*.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., Rebeyrolle, J. et Tanguy, L. (2010). From cues to markers: a corpus-driven approach to complex discourse organisation markers. *Linguistic and Psycholinguistic Approaches to Text Structuring. Sarda, L., Carter Thomas, C. et Fagard, B. (coord.). A paraître*.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press, Oxford.

Ibekwe-SanJuan, F. (2005). Inclusion lexicale et proximité sémantique entre termes. *In Actes des 6e rencontres Terminologie et Intelligence Artificielle (TIA)*, pages 45–57.

Ibrahim, A., Katz, B. et Lin, J. (2003). Extracting structural paraphrases from aligned monolingual corpora. *In Second international workshop on Paraphrasing*, pages 57–64, Morristown, NJ, USA.

Jacquemin, C. (1997). *Variation terminologique: Reconnaissance et acquisition automatique de termes et de leurs variantes, 1997.* Habilitation à diriger des recherches. Université de Nantes.

Jacquemin, C., Klavans, J. L. et Tzoukermann, E. (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. *In Proceedings of ACL-EACL*, pages 24–31, Madrid.

Jacquemin, C. et Tzoukermann, E. (1999). NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon, and Syntax. *In* Strzalkowski, T., éditeur : *Natural Language Information Retrieval*, pages 25–74. Kluwer, Boston, MA.

Jones, S. (2002). *Antonymy - A corpus-based Perspective.* Routledge.

Justeson, J. et Katz, S. (1991). Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, 17(1):1–19.

Katz, B. et Lin, J. (2003). Selectively using relations to improve precision in question answering. *In Proceedings of the EACL Workshop on Natural Language Processing for Question Answering*, pages 43–50, Budapest.

Kilgarriff, A. et Yallop, C. (2000). What's in a thesaurus. *In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pages 1371–1379, Athènes. Citeseer.

Korhonen, A., Krymolowski, Y. et Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Gênes, Italie. Citeseer.

Kupść, A. (2007). Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. *In Actes des 14èmes journées sur le Traitement Automatique des Langues Naturelles (TALN)*, Toulouse.

Kupść, A. (2008). Adjectives in TreeLex. *In M. Klopotek, A. Przepiórkowski, S. Wierzchoń et K. Trojanowski (eds.), 16th International Conference Intelligent Information Systems*, pages 287–296, Zakopane, Pologne.

Laignelet, M. (2009). *Analyse discursive pour le repérage automatique de segments obsolescents dans des documents encyclopédiques.* Thèse de doctorat, Université Toulouse 2 - Le Mirail.

Lapata, M. (2001). A corpus-based account of regular polysemy: The case of context-sensitive adjectives. *In Proceedings of the 2nd Meeting of the NAACL*, pages 63–70, Pittsburgh, Pennsylvania.

Laurent, D., Nègre, S. et Séguéla, P. (2006). QRISTAL, le QR à l'épreuve du public. *Traitement automatique des langues (TAL)*, 46(3):41–70.

Le Draoulec, A. et Fabre, C. (2006). La dimension événementielle du syntagme nominal dans la structure avant+ SN. *Cahiers de lexicologie*, 89:47–74.

Leeman, D. (1999). "Dans un juron, il sauta sur ses pistolets" : aspects de la polysémie de la préposition dans. *Revue de Sémantique et Pragmatique*, 6:71–88.

Legallois, D. (2005). Du bon usage des expressions idiomatiques dans l'argumentation de deux modèles anglo-saxons. *Cahiers de l'Institut de linguistique de Louvain*, 31(2):109.

Legallois, D. et François, J. (2006). Autour des grammaires de constructions et de patterns. *Cahiers du CRISCO*, 21.

Léglise, I. (1997). Intervention linguistique: théorie, pratique et intérêt dans le cadre de l'analyse de l'activité. *Linx*, 37:169–182.

Léglise, I. (2000). Quand les linguistes interviennent: écueils et enjeux. *Revue française de linguistique appliquée*, IV:5–13.

L'Homme, M.-C. (2002). What can verbs and adjectives tell us about terms? *In Proceedings of Terminology and Knowledge Engineering (TKE)*, pages 28–30, Nancy.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. *In Proceedings of the Association of Computational Linguistics Conference (ACL)*, volume 36, pages 768–774, Montréal.

Lin, D. (1998b). An information-theoretic definition of similarity. *In Proceedings of the Fifteenth International Conference on Machine learning (ICML)*, Madison, Wisconsin.

Lin, D. et Pantel, P. (2002). Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(04):343–360.

Lin, D., Zhao, S., Qin, L. et Zhou, M. (2003). Identifying synonyms among distributionally similar words. *In Proceedings of the International joint Conference on Artificial Intelligence*, volume 18, pages 1492–1493.

Lobanova, A., Van Der Kleij, T. et Spenader, J. (2010). Defining Antonymy: A Corpus-based Study of Opposites by Lexico-syntactic Patterns. *International Journal of Lexicography*, 23(1):19–53.

Manning, C. (2003). Probabilistic syntax. *In* R. Bod, J. H. et Jannedy, S., éditeurs : *Probabilistic linguistics*, pages 289–341. Cambridge, MA: MIT Press.

McEnery, T., Xiao, R. et Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge, New-York.

Mel'čuk, I. (1988). Paraphrase et lexique dans la théorie linguistique sens-texte. *Lexique*, 6:13–54.

Mel'čuk, I. et Polguère, A. (2007). *Lexique actif du français: l'apprentissage du vocabulaire fondé sur 20000 dérivations sémantiques et collocations du français.* De Boeck, Bruxelles.

Merlo, P. et Ferrer, E. (2006). The notion of argument in Prepositional Phrase attachment. *Computational Linguistics*, 32(2):341–378.

Merlo, P. et Leybold, M. (2001). Automatic Distinction of Arguments and Modifiers: the Case of Prepositional Phrases. *In Workshop on Computational Language Learning (Connl 2001)*, Toulouse.

Messiant, C., Korhonen, A. et Poibeau, T. (2008). Lexschem: A large subcategorization lexicon for french verbs. *In Actes de la 6ème conférence LREC, Language Resources and Evaluation Conference*, Marrakech, Maroc.

Milićević, J. (2003). *Modélisation sémantique, lexicale et syntaxique de la paraphrase.* Thèse de doctorat, Département de linguistique, Université de Montréal.

Miller, P. (1998). Compléments et circonstants: une distinction syntaxique ou sémantique. *Cycnos*, 15:91–103.

Milner, J. (1989). *Introduction à une science du langage.* Seuil.

Mirkin, S., Dagan, I. et Geffet, M. (2006). Integrating pattern-based and distributional similarity methods for lexical entailment acquisition. *In Proceedings of the COLING/ACL on Main conference poster sessions*, page 586. Association for Computational Linguistics.

Morlane-Hondère, F. (2008). Combiner proximité distributionnelle et patrons linguistiques pour repérer l'antonymie. mémoire de Master 1 en sciences du langage, Université Toulouse 2 - Le Mirail.

Morris, J. et Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Morris, J. et Hirst, G. (2004). Non-classical Lexical Semantic Relations. *In Proceedings of the Workshop on Computational Lexical Semantics, Human Language Technology Conference of the NAACL*, pages 46–51, Boston.

Mortureux, M.-F. (1993). Paradigmes désignationnels. *Semen*, 8:123–141.

Mothe, J. et Tanguy, L. (2005). Linguistic features to predict query difficulty - a case study on previous trec campaigns. *In SIGIR workshop on Predicting Query Difficulty - Methods and Applications*, pages 7–10.

Murphy, L. (2003). *Semantic Relations and the Lexicon. Antonymy, Synonymy, and Other Paradigms.* Cambridge University Press.

Murphy, M. (2006). Antonyms as lexical constructions: or, why paradigmatic construction is not an oxymoron. *Constructions*, 1:1–8.

Nazarenko, A. (2005). Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel. *A. Condamines (ed.), Sémantique et Corpus (Traité IC2, série Cognition et traitement de l'information), Hermès*, 31:211–244.

Neelameghan, A. (2001). *Lateral relationships in multicultural, multilingual databases in the spiritual and religious domains: The OM Information Service*, pages 185–198. Springer.

Nespoulous, J.-L. et Virbel, J. (2003). Vers une révision de la notion de lexicalisation - contribution à une vision dynamique du lexique mental : 'stock' lexical, catégories vs 'réseau' lexico-sémantique. *Revue d'Intelligence Artificielle*, 17(5-6):747–760.

Pène, S., Borzeix, A. et Fraenkel, B. (2001). *Le langage dans les organisations: une nouvelle donne*. L'Harmattan.

Péry-Woodley, M. (2000). *Une pragmatique à fleur de texte: approche en corpus de l'organisation textuelle*. Habilitation à diriger des recherches. Université Toulouse 2 - Le Mirail.

Péry-Woodley, M. et Scott, D. (2006). Computational Approaches to Discourse and Document Processing. *T.A.L.*, 47(2):7–19.

Picton, A., Fabre, C. et Bourigault, D. (2008). Méthodes linguistiques pour l'expansion de requêtes. Une expérience basée sur l'utilisation du voisinage distributionnel. *Revue Française de Linguistique Appliquée*, XIII(1):83–96.

Poibeau, T. et Condamines, A. (2008). Linguistique et accès automatisé à l'information: un bilan. *Revue française de linguistique appliquée*, 13(1):5–8.

Poibeau, T. et Messiant, C. (2008). Do we still need gold standard for evaluation. *In Proceedings of the Language Resources and Evaluation Conference (LREC), Marrakech*.

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

Römer, U. (2005). *Progressives, patterns, pedagogy: a corpus-driven approach to English progressive forms, functions, contexts, and didactics*. John Benjamins Publishing Co.

Rosario, B. et Hearst, M. (2001). Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. *In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–90.

Sagot, B., Fort, K. et Venant, F. (2009). Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes du français. *Linguisticæ Investigationes*, 32(2):305–315.

Spärck-Jones, K. (2007). Computational linguistics: what about the linguistics? *Computational linguistics*, 33(3):437–441.

Stede, M. (1996). Lexical paraphrases in multilingual sentence generation. *Machine Translation*, 11(1):75–107.

Stefanowitsch, A. et Gries, S. (2009). Corpora and grammar. *A. L "udeling und M. Kyt "o (eds), Corpus Linguistics. An International Handbook*, 29:933–952.

Tanguy, L. et Hathout, N. (2007). *Perl pour les linguistes. Programmes en Perl pour exploiter les données langagières.* Collection TIC et sciences cognitives, Lavoisier.

Tognini Bonelli, E. (2001). *Corpus Linguistics at Work.* John Benjamins, Amsterdam.

Tudhope, D., Alani, H. et Jones, C. (2001). Augmenting thesaurus relationships: possibilities for retrieval. *Journal of Digital information*, 1(8):1.

Turney, P. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Tutin, A. (2007). Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. *Actes de TALN 2007. Communications affichées*, pages 283–292.

Tutunjian, D. et Boland, J. (2008). Do We Need a Distinction between Arguments and Adjuncts? Evidence from Psycholinguistic Studies of Comprehension. *Language and Linguistics Compass*, 2(4):631–646.

Vaguer, C. (2005). Pourquoi sombre-t-on dans le malheur?. Étude de constructions verbales «V dans Némotion». *Lidil. Revue de linguistique et de didactique des langues*, (32).

Valette, M. (2010). Approche textuelle du lexique. *Texto ! [En ligne], URL : http://www.revue-texto.net/index.php?id=2613.*

Van der Plas, L. (2008). *Automatic lexico-semantic acquisition for question answering.* Thèse de doctorat, PhD, Rijksuniversiteit Groningen.

Vergely, P. (2004). *Analyse linguistique de l'expression du dysfonctionnement technique : le cas des échanges entre chefs de salle et maintenance opérationnelle dans la Navigation Aérienne.* Thèse de doctorat, Université Toulouse 2 - Le Mirail.

Vergely, P., Condamines, A., Fabre, C., Josselin-Leray, A., Rebeyrolle, J. et Tanguy, L. (2009). Analyse linguistique des disparités dans les échanges médecin-patient. *Actes du colloque Actes éducatifs et soins : entre éthique et gouvernance, atelier "l'acte de soin approché par les aspects langagiers".*

Vossen, P. (1998). *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

Vossen, P. (2008). Linguistic knowledge for more precision, richer answers and flexible systems. *Revue française de linguistique appliquée*, XIII(1):24–39.

Wanner, L. (2004). Towards automatic fine-grained semantic classification of verb-noun collocations. *Natural Language Engineering*, 10(2):95–143.

Wintner, S. (2009). What science underlies natural language engineering? *Computational Linguistics*, 35(4):641–644.

Xu, J. et Croft, W. (1998). Corpus-based stemming using cooccurrence of word variants. *In ACM Transactions on Information Systems (TOIS), Computer Science Technical Report TR96-67*, volume 16, pages 61–81.

Zaenen, A. (2006). Mark-up barking up the wrong tree. *Computational Linguistics*, 32(4):577–580.

Zweigenbaum, P., Grau, B., Ligozat, A., Robba, I., Rosset, S., Tannier, X., Vilnat, A. et Bellot, P. (2008). Apports de la linguistique dans les systèmes de recherche d'informations précises. *(Revue Française de Linguistique Appliquée)*, XIII:41–62.